האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

SELECT SETS: RANK AND FILE

by

ABBA M. KRIEGER, MOSHE POLLAK
and ESTER SAMUEL-CAHN

Discussion Paper # 388 March 2005

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Select Sets: Rank and File

Abba M. Krieger*
University of Pennsylvania

Moshe Pollak†
Hebrew University

Ester Samuel-Cahn‡
Hebrew University

Abstract

In many situations, the decision maker observes items in sequence and needs to determine whether or not to retain a particular item immediately after it is observed. Any decision rule creates a set of items that are selected. We consider situations where the available information is the rank of a present observation relative to its predecessors. Certain "natural" selection rules are investigated. Theoretical and Monte Carlo results are presented pertaining to the evolution of the number of items selected, measures of their quality and the time it would take to amass a group of a given size. A comparison between rules is made, and guidelines to the choice of good procedures are offered.

Key words: Selection rules; Ranks; Nonparametrics; Sequential observations; Asymptotics
AMS 2000 Classification: Primary: 62L99 Secondary: 62F07, 60F15

*Abba M. Krieger is Robert Steinberg Professor and Chair, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

†Moshe Pollak is Marcy Bogen Professor of Statistics, Department of Statistics, Hebrew University of Jerusalem Israel. The work of Abba M. Krieger and Moshe Pollak was supported by funds from the Marcy Bogen Chair.

‡Ester Samuel-Cahn is the Mr. and Mrs. Hock Professor of Statistics, Department of Statistics, Hebrew University, Jerusalem Israel. Ester Samuel-Cahn's work was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 467/04).

1 Introduction

Consider a newly formed company in immediate need of a cadre of salespeople. Applicants are interviewed sequentially, each receiving a score. The decision whether or not to hire a candidate must be made on the spot, without possibility of getting back to someone who has been let go. What would constitute a reasonable hiring policy?

The issues involved in formulating a policy are the quality and quantity of those hired and the speed at which positions are filled. Also of import is whether or not the horizon of the candidate pool is finite. The scenario we envision is one where applicants arrive in random order and their scores are independent and identically distributed, but nothing is otherwise known about their distribution, so that as applicants are interviewed information about the pool of candidates is being gathered. Heuristically, quality of the hired staff and speed of filling positions are in conflict with each other. A policy of hiring everyone answers the need for speed, but the quality will be average. Towards the other extreme, declining to hire anyone if he or she is not better than all of those observed previously will produce a high-quality group of salespeople, but its rate of growth will be very slow. In this paper we study certain policies that compromise between the two objectives.

The procedure that accepts the first applicant and subsequently accepts only candidates who are better than all those observed previously is a well-studied policy (cf. Arnold, Balakrishnan and Nagaraja, 1998, and Resnick, 1987).

Preater (2000) studied a method that prescribes the employment of the first applicant and subsequently engages only those who, if hired, would increase the average score of those retained. Preater assumed that the scores are exponentially distributed, and derived the asymptotic growth and distribution of the average score after n observations have been retained, as $n \to \infty$. Selection rules with known distribution of the items and inspection cost are considered in Preater(1994).

Other problems that have a similar flavor are variations of the secretary problem (cf. Gilbert and Mosteller, 1966), where one samples sequentially from a finite pool until one or several are retained, after which sampling ceases, with the objective being the maximization of the probability of retaining the best in the pool.

In this paper, we study sequential rules that are based on the ranks of the observations. At every stage, we (re-) rank the observed values from the best to the worst, so that the best has rank 1. We consider procedures that retain the n^{th} observation if its rank is low enough relative to the ranks of the previously retained observations.

For the sake of illustration, consider the median rule, which prescribes the retention of the n^{th} observation if and only if its rank is lower than the median rank of the observations retained previously (i.e., the median of the retained group will be improved). Regarding the speed at which observations are retained, let L_n be the number of items that are retained after n items are observed. We show that the expected number $E(L_n)$ of observations

retained after n have been observed is of order $n^{1/2}$ and that $L_n/n^{1/2}$ converges almost surely to a non-degenerate random variable. Regarding the quality of the retained observations, we show that at least half of the observations retained are the very best of all observed heretofore, and that the expectation of the average rank of the retained observations is of order $n^{1/2}\log n$ (implying that almost all of the retained observations are very good). We show by Monte Carlo that, if the prospective pool has size 10,000, the median rule retains 118 on the average, and approximately 70% of those retained are among the top 1% of the pool, and only 4% don't make it to the top decile.

The paper is organized as follows. In Section 2, we introduce a general class of rules that are characterized by a criterion that ensures that the probability that item n is retained is a simple function of the number of items L_n that have already been retained. In Section 3 we specialize and consider rules that retain an item if it is among the best p percent of the items already retained. We show that $E(L_n)/n^p$ converges. In Section 4 we show that L_n/n^p converges almost surely to a non-degenerate random variable. In Section 5 we find the order of the expected value of the average rank of the observations retained by the p-percentile rule and that suitably normalized, the average converges a.s. to a nondegenerate random variable. Section 6 is devoted to a Monte Carlo study. We end with Remarks and Conclusions in Section 7.

2 A class of selection rules

As stated in the previous section, our focus in this article is on selection rules based on ranks. In this section we introduce a class of selection rules that retain an observation if its rank is "low enough", where the threshold of "low enough" is determined solely by the size of the set of observations already retained. The rationale for this has to do with the tradeoff between the quality of retained observations and the speed of their accumulation. Heuristically, the more observations retained, the slower one would be about retaining further observations, so the size of the retained set should be a factor in the selection rule. On the other hand, one's evaluation of the quality of an observation depends on all past observations, not only on those retained so far, and one's expectations regarding future observations is the same irrespective of the quality of those already retained. Therefore, there is good reason to require a selection rule to depend only on the size of the retained set of observations and the rank of the present observation. (As for the desire to "improve" the set of retained observations, heuristically, the quality of the retained set is correlated with its size, so at least qualitatively "improvement" is implicit in retained set size. We examine this more formally in Theorem 1.)

Formally, let $X_1, X_2, \ldots$ be a sequence of observations so that any ordering of the first n observations is equally likely. A sufficient condition is that the random variables be exchangeable. A special case that satisfies this assumption is when we have independent

and identically distributed (i.i.d.) random variables from a continuous distribution. Let S_n be the set of indices of the retained $X's$. Let L_n be the size of S_n . Let R_i^n be the rank of the i^{th} observation from among $X_1, \cdots, X_n$, i.e., $R_i^n = \sum_{j=1}^n I\{X_i \leq X_j\}$ where $I\{A\}$ is the indicator function of A. Thus, R_n^n is the rank of X_n within the set $\{X_i\}_{i=1}^n$, where without loss of generality we assume that "better" is equivalent to "larger" so that rank 1 is given to the largest observation, rank 2 to the second largest, etc. A selection rule of the type we study is defined by a function $r()$ on the integers such that the observation X_n will be retained if and only if $R_n^n \leq r(L_{n-1})$. In this article, we assume that the first observation is kept always.

Another feature of a reasonable selection procedure is to require that the function $r()$ be locally sub-diagonal; i.e. $r(a+1) \leq r(a)+1$. Again, the rationale for this has to do with the tradeoff between the quality of retained observations and the speed of their accumulation. (To see this, suppose a observations have been retained after n have been observed. The rank of the next retained observation will not exceed $r(a)$. The rank of the succeeding retained observation will not exceed $r(a+1)$. If $r(a+1) > r(a)+1$, it would mean that after having retained $a+1$ observations, one would be willing to settle for an observation of lower quality than the acceptance level after having retained a observations. Although that may be reasonable in a case that a quota has to be filled and the pool of applicants is finite, that is not the case we regard here.)

We summarize the above in the following definition.

Definition 1 *A locally sub-diagonal rank selection scheme (LsD) is a rule determined by a function $r()$ with the following properties:*

i) r is non-decreasing.

ii) $r(0) = 1$ and $L_0 = 0$.

iii) r is locally sub-diagonal; i.e. $r(a+1) \leq r(a)+1$.

iv) For $n \geq 1$, X_n is retained if and only if $R_n^n \leq r(L_{n-1})$. (This means that the first observation is retained.)

This class contains many rules that make heuristic sense. For instance, the median rule is a LsD rule with $r(1) = r(2) = 1$, $r(3) = r(4) = 2$, and generally $r(2j-1) = r(2j) = j$. A class of LsD rules is "k-record rules". For a fixed value k, let $r(j) = \min\{j+1, k\}$. For $k = 1$, this is the classical record rule, where an element is retained if and only if it is better than all previous observations. "k-record rules" have been extensively studied (cf. Leadbetter, Lindgren and Rootzen, 1983, or Resnick, 1987, and many subsequent papers).

The following observation is trivial for "k-record rules" but is true for any LsD rule. It attests to the high quality of the set of observations retainable by a LsD rule.

Let $N \geq 1$ be any integer either predetermined or random. For example, N can be a stopping rule. A special case of interest is inverse sampling, where the objective is to

collect a group of some fixed size m, so that

$$N = \inf\{n : L_n = m\}.$$

Theorem 1 *Consider a LsD rule defined by r(). The $r(L_N)$ best observations among $X_1, X_2, ..., X_N$ belong to S_N.*

Proof: Let X_m be the t^{th} best observation among $X_1, X_2, ..., X_N$ with $t \leq r(L_N)$. Let a be the number of observations among $X_1, X_2, ..., X_{m-1}$ that are better than X_m and let b be the number of observations among $X_{m+1}, ..., X_N$ that are better than X_m . Clearly, $a + b = t - 1$.

If $m \notin S_N$, then the next item retained after iteration m must be better than X_m . This implies that $L_N \leq L_{m-1} + b$. Hence

$$r(L_N) \leq r(L_{m-1} + b) \leq r(L_{m-1}) + b. \tag{1}$$

But $m \notin S_N$ implies that $a + 1 > r(L_{m-1})$. Since by assumption $t \leq r(L_N)$,

$$a + b + 1 = t \leq r(L_N) \leq r(L_{m-1}) + b \tag{2}$$

so that $a + 1 \leq r(L_{m-1})$, which contradicts the inequality two lines above. ∎

Remark: Because of Theorem 1, implicit in the definition of a LsD rule is that it "improves" the retained set. For example, when applying the median rule, the median of the retained set gets better, something that is not transparent when regarding the median rule via its LsD definition. Theorem 1 is a more formal presentation of the heuristic stated in the beginning of this section, that the quality of the retained set is correlated with its size, and a LsD rule embodies all three heuristics: i) the larger the retained set, the slower one goes about retaining more observations; ii) perception of quality is founded on all previous observations; iii) one only retains items that "improve" the retained set. Theorem 1 means that there is no contradiction between the third heuristic and selection based on the size of the retained set only.

A natural representation of the quality of the group of observations kept (when retention is by ranks) is the average rank of the observations retained. Denote by Q_n the sum of the ranks of the retained set after n observations have been made, so that the average rank A_n is Q_n/L_n.

Assertion Let $\mathcal{F}_n$ be the $\sigma - field$ formed by the random variables $X_1, \cdots, X_n$. Let $Q_n = \sum_{i \in S_n} R_i^n$. The conditional expected behavior of the quantities L_{n+1}, Q_{n+1} and the average rank A_{n+1} given the past, in terms of the corresponding quantities for n, results in

i) $E(L_{n+1}|\mathcal{F}_n) = L_n + \frac{r(L_n)}{n+1}$

ii) $E(Q_{n+1}|\mathcal{F}_n) = \frac{n+2}{n+1}Q_n + \frac{r(L_n)(r(L_n)+1)}{2(n+1)}$

iii) $E\left(A_{n+1}|\,\mathcal{F}_n\right) = A_n\left(1 + \frac{1+L_n-r(L_n)}{(n+1)(L_n+1)}\right) + \frac{r(L_n)(r(L_n)-1)}{2(n+1)L_n}$

Proof: i) $L_{n+1}|L_n = L_n + B\left(\frac{r(L_n)}{n+1}\right)$ where $B(x)$ is a Bernoulli random variable with probability x. Hence (i) follows by taking conditional expectations on both sides.

ii) $\{Q_n\}$ is a non-decreasing sequence. Its growth can be described as follows. If X_{n+1} is retained, and it has rank k among the items retained, then X_{n+1} adds k to the sum of the ranks and one for each observation that is inferior to it. Hence, $Q_{n+1} = Q_n + k + [L_n - (k-1)] = Q_n + L_n + 1$. When X_{n+1} is not retained, then the rank of each of the lower-quality retained observations can increase (by 1, if X_{n+1} has lower rank). Note that the distribution of the rank of X_{n+1} (conditional on $\mathcal{F}_n$ and its not being retained) is uniform over $r(L_n)+1, \cdots, n+1$. Therefore, letting S_n = the retained set after n observations have been made, for $n \geq 1$

$$\begin{aligned}
E\left(Q_{n+1}|\mathcal{F}_n\right) &= (Q_n + L_n + 1)\frac{r(L_n)}{n+1} \\
&\quad + \frac{n+1-r(L_n)}{n+1}\left[Q_n + E\left(\sum_{\{i\in S_n\}} I\left\{R_{n+1}^{n+1} < R_i^{n+1}\right\}|\mathcal{F}_n, R_{n+1}^{n+1} > r(L_n)\right)\right] \\
&= Q_n + (L_n+1)\frac{r(L_n)}{n+1} + \frac{Q_n - r(L_n)(r(L_n)+1)/2 - r(L_n)(L_n - r(L_n))}{n+1} \\
&= \frac{n+2}{n+1}Q_n + \frac{r(L_n)(r(L_n)+1)}{2(n+1)}.
\end{aligned}$$

iii) If X_{n+1} is retained then $L_{n+1} = L_n + 1$ and if X_{n+1} is not retained $L_{n+1} = L_n$. Therefore, the same argument as in the proof of ii) leads to

$$\begin{aligned}
E\left(A_{n+1}|\,\mathcal{F}_n\right) &= \frac{Q_n + L_n + 1}{L_n + 1}\cdot\frac{r(L_n)}{n+1} \\
&\quad + \frac{n+1-r(L_n)}{n+1}\left[\frac{Q_n + E\left(\sum_{\{i\in S_n\}} I\left\{R_{n+1}^{n+1} < R_i^{n+1}\right\}|\mathcal{F}_n, R_{n+1}^{n+1} > r(L_n)\right)}{L_n}\right] \\
&= A_n\left[\frac{L_n}{L_n+1}\cdot\frac{r(L_n)}{n+1}\right] + \frac{r(L_n)}{n+1} \\
&\quad + \frac{n+1-r(L_n)}{(n+1)L_n}\left[Q_n + E\left[\frac{\sum_{\{i\in S_n, R_{n+1}^{n+1} < R_i^{n+1}\}}(R_i^{n+1} - r(L_n))}{n+1-r(L_n)}\middle|\,\mathcal{F}_n\right]\right] \\
&= A_n\left[\frac{r(L_n)L_n}{(n+1)(L_n+1)} + \frac{n+1-r(L_n)}{n+1}\right] \\
&\quad + \frac{r(L_n)}{n+1} + \frac{1}{(n+1)L_n}\left[Q_n - \frac{r(L_n)(r(L_n)+1)}{2} - r(L_n)(L_n - r(L_n))\right]
\end{aligned}$$

$$= A_n \left[1 + \frac{1 + L_n - r(L_n)}{(n+1)(L_n+1)}\right] + \frac{r(L_n)(r(L_n)-1)/2}{(n+1)L_n}. \quad \blacksquare \tag{3}$$

3 Percentile rules

In the following sections, we consider rules that retain items if the item is among the best p-percent among those items that have already been retained.

Definition: A p-percentile rule, for p fixed ($0 < p \leq 1$) is a LsD rule with $r(k) = \lceil pk \rceil$ for $k \geq 1$, where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x. Thus, the n^{th} item is retained if and only if its rank satisfies $R_n^n \leq \lceil pL_{n-1} \rceil$.

To see that the p-percentile rule is a LsD rule, note that $\lceil p(a+1) \rceil = \lceil pa + p \rceil \leq \lceil pa + 1 \rceil = \lceil pa \rceil + 1$.

Remark: Note that the p-percentile rule is meaningful even when $p = 1$. In that case, the first observation is kept. The second is kept if it is better than the first observation. In general, an item is kept if this is better than the worst item that has already been retained. It is easy to see that when $p = 1$, $E(L_n|L_{n-1}) = L_{n-1} + L_{n-1}/n$. It is straightforward to show that $E(L_n) = \frac{n+1}{2}$. Hence $E(L_n)/n \to 1/2$. Also, since $E(L_n|L_{n-1}) = \frac{n+1}{n} L_{n-1}$, it follows that $L_n/(n+1)$ is a bounded positive martingale, and therefore converges almost surely. Since the worst item that has already been retained is obviously X_1, it follows that L_n/n is asymptotically $U(0,1)$.

4 Results for the Number of Retained Items

In this section, we study the behavior of the number of items that are retained after n items are observed, L_n, for $p-$percentile rules. It turns out that L_n is of order n^p. Hence we consider the quantity L_n/n^p. We first show that the expectation of this quantity converges to a finite limit. We then show that this quantity converges almost surely to a non-degenerate random variable.

The first result we present is that $E(L_n)/n^p \to c_p$ as $n \to \infty$. For example, this results says that the rule that retains items if they are superior to the median of all items already retained, will be keeping on the order of $\sqrt{n}$ items on the average. The constant c_p depends on

$$d_n \equiv E(\lceil pL_n \rceil - pL_n). \tag{4}$$

The relationship between c_p and $d_1, d_2, ...$ is complicated because it depends on all of the

d_j. It seems impossible to determine c_p analytically, except for $p = 1$, as done in the remark above.

The result, however, only requires that we show that d_n is bounded away from zero. This result is intuitive. For the median rule ($p = 1/2$), d_n is simply $P(L_n \text{ is odd})/2$. Logically, we would expect (it turns out to be justified by empirical analysis) that $P(L_n \text{ is odd}) \to 1/2$ as $n \to \infty$. This is not easy to prove. Similarly, if $p = 1/4$, then $\lceil pL_n \rceil - pL_n$ is either 0, 3/4, 1/2, or 1/4 depending on whether $L_n(mod4)$ is j for $j = 0$,1 ,2 or 3 respectively. Since logically each of the four cases should be equally likely (again this appears to be the case by computer analysis), we would anticipate that $d_n \to 3/8$. We conjecture that if p is an irrational number, then $\lceil L_n p \rceil - L_n p$ converges to $U(0, 1)$ which implies that $d_n \to 1/2$. The actual value of d_n is hard to determine. But it is sufficient to show that d_n is bounded away from zero. Specifically,

Lemma 1 *Let $0 < p < 1$ be fixed, and $\epsilon = \epsilon_p = min\{\frac{p}{2}, \frac{1-p}{2}\}$. Then $d_n \geq \epsilon/3$ for all n.*

Proof: Let $S_\epsilon = \{j \mid \lceil pj \rceil - pj \leq \varepsilon\}$. Note that if $j \in S_\epsilon$, then

- $j-1 \notin S_\epsilon$. This follows since $\epsilon+p < 1$, thus $\lceil p(j-1) \rceil = \lceil pj \rceil$. But then $\lceil p(j-1) \rceil - p(j-1) = \lceil pj \rceil - pj + p \geq p > \epsilon$.
- $j+1 \notin S_\epsilon$ This follows since $p-\epsilon > 0$, thus $\lceil p(j+1) \rceil = \lceil pj \rceil + 1$. Hence $\lceil p(j+1) \rceil - p(j+1) = \lceil pj \rceil - pj + 1 - p > \epsilon$.

We will show that for all $n \geq 2$ and all $j = 1, 2, \ldots$

$$P(L_n = j+1) + P(L_n = j-1) - P(L_n = j) \geq 0. \tag{5}$$

This will yield the lemma since clearly (5) implies $\sum_{j \in S_\epsilon} P(L_n = j) \leq 2 \sum_{j \notin S_\epsilon} P(L_n = j)$, which in turn implies that $\sum_{j \notin S_\epsilon} P(L_n = j) \geq 1/3$ so $d_n \geq \epsilon/3$. Note that (5) is trivial for $j > n$.

We prove (5) by induction. For $n = 2$ and all $0 < p < 1$ we have $P(L_2 = 1) = P(L_2 = 2) = 1/2$. Thus (5) holds for $j = 1, 2$ and $n = 2$.

Now assume (5) holds for $2, 3, \ldots, n-1$. We shall show it holds for n. Consider first the values of j for which

$$2\lceil pj \rceil / n \leq 1. \tag{6}$$

Clearly

$$\begin{aligned}
P(L_n = j-1) &\geq (1 - \lceil p(j-1) \rceil / n) P(L_{n-1} = j-1) && (7)\\
P(L_n = j) &= (\lceil p(j-1) \rceil / n) P(L_{n-1} = j-1) + (1 - \lceil pj \rceil / n) P(L_{n-1} = j) && (8)\\
P(L_n = j+1) &= (\lceil pj \rceil / n) P(L_{n-1} = j) + (1 - \lceil p(j+1) \rceil / n) P(L_{n-1} = j+1). && (9)
\end{aligned}$$

Thus

$$\begin{aligned} & P(L_n = j+1) + P(L_n = j-1) - P(L_n = j) \\ \geq\ & P(L_{n-1} = j+1) + P(L_{n-1} = j-1) - P(L_{n-1} = j) + 2(\lceil pj \rceil / n) P(L_{n-1} = j) \\ -\ & 2(\lceil p(j-1) \rceil / n) P(L_{n-1} = j-1) - (\lceil p(j+1) \rceil / n) P(L_{n-1} = j+1). \end{aligned} \tag{10}$$

However, $\lceil p(j-1) \rceil \leq \lceil pj \rceil$ and $\lceil p(j+1) \rceil \leq 2\lceil pj \rceil$ as $\lceil pj \rceil \geq 1$. Hence, the right hand side of (10) is greater than or equal to

$$(1 - 2\lceil pj \rceil / n)[P(L_{n-1} = j+1) + P(L_{n-1} = j-1) - P(L_{n-1} = j)] \geq 0 \tag{11}$$

where the last inequality in (11) follows from (6) and the induction hypothesis.

Now consider values of j (if such exist) for which

$$2\lceil pj \rceil / n > 1. \tag{12}$$

Then clearly $j > 1$. Replace (7) by

$$P(L_n = j-1) = (1 - \lceil p(j-1) \rceil / n) P(L_{n-1} = j-1) + (\lceil p(j-2) \rceil / n) P(L_{n-1} = j-2) \tag{13}$$

and replace (9) by

$$P(L_n = j+1) \geq (\lceil pj \rceil / n) P(L_{n-1} = j). \tag{14}$$

Then by (13), (8), and (14), it follows that

$$\begin{aligned} P(L_n = j+1) + P(L_n = j-1) - P(L_n = j) \geq (2\lceil pj \rceil / n - 1) P(L_{n-1} = j) \\ + (\lceil p(j-2) \rceil / n) P(L_{n-1} = j-2) - (2\lceil p(j-1) \rceil / n - 1) P(L_{n-1} = j-1). \end{aligned} \tag{15}$$

If $2\lceil p(j-1) \rceil / n \leq 1$ then by (12) clearly the value in the right hand side of (15) is non-negative. If

$$2\lceil p(j-1) \rceil / n - 1 > 0 \tag{16}$$

we shall show that (16) implies

$$\lceil p(j-2) \rceil / n \geq 2\lceil p(j-1) \rceil / n - 1 \tag{17}$$

so that the right hand side of (15) is greater or equal to

$$(2\lceil p(j-2) \rceil / n - 1)[P(L_{n-1} = j) + P(L_{n-1} = j-2) - P(L_{n-1} = j-1)] \geq 0 \tag{18}$$

where the last inequality follows from (16) and the induction hypothesis. To see (17) note that $\lceil p(j-2) \rceil \geq \lceil p(j-1) \rceil - 1$. Thus (17) will follow if we show that $(\lceil p(j-1) - 1)/n \geq 2\lceil p(j-1) \rceil / n - 1$ which is equivalent to

$$n - 1 \geq \lceil p(j-1) \rceil. \tag{19}$$

Since $j \leq n$ are the only values of interest, we have $j-1 \leq n-1$, for which (19) clearly holds. ∎

We now turn to the main result of showing that the average number of items that are retained is of order n^p. From Assertion i) at the end of Section 2,

$$E(L_n|L_{n-1}) = L_{n-1} + \lceil pL_{n-1} \rceil /n. \tag{20}$$

Hence,

$$E(L_n|L_{n-1}) = L_{n-1} + pL_{n-1}/n + (\lceil pL_{n-1} \rceil - pL_{n-1})/n \tag{21}$$

Let $M_n = E(L_n)$. Then

$$M_n = M_{n-1}(1+p/n) + d_{n-1}/n. \tag{22}$$

We are now prepared to state and prove the theorem.

Theorem 2 *Let* $0 < p \leq 1$. $M_n/n^p \to c_p$ *as* $n \to \infty$ *with* $0 < c_p < \infty$, *where* $M_n = E(L_n)$.

Proof. By the Remark in Section 3, $c_1 = 1/2$. For $0 < p < 1$, let $T_n = M_n/n^p$. From (22) we have that

$$T_n = ((n-1)/n)^p(1+p/n)T_{n-1} + d_{n-1}/n^{1+p}. \tag{23}$$

The key to the proof is showing that $\Delta_n \equiv T_n - T_{n-1}$ eventually becomes positive and remains positive. Since $T_n = \sum_{j=1}^n \Delta_j$ with $T_0 \equiv 0$ and T_n will be shown to be bounded, it follows that T_n converges.

By definition of Δ_j, and (23)

$$\Delta_j = b_j T_{j-1} + d_{j-1}/j^{1+p} \tag{24}$$

where $b_j = ((j-1)/j)^p(1+p/j) - 1$.

The basis of the proof is in the result that $b_j < 0$ and increases to 0 as $j \to \infty$. This is a straightforward calculus argument.

Let $x = 1/j$ and $f(x) = (1-x)^p(1+px) - 1$. Thus, $f(0) = 0$. Also

$$\begin{aligned} f'(x) &= p(1-x)^p - (1+px)p(1-x)^{p-1} \\ &= p(1-x)^{p-1}[1-x-1-px] \\ &= -p(1+p)x(1-x)^{p-1} < 0. \end{aligned}$$

Since $x = 1/j$, this shows that b_j increases to zero as $j \to \infty$.

From $b_j < 0$ and (24) it follows that

$$T_n \leq 1 + \sum_{j=2}^n \frac{1}{j^{1+p}} \leq 1 + \int_{x=1}^n (1/x)^{1+p} dx \leq 2/p \tag{25}$$

so T_n is bounded. To show that Δ_n is eventually non-negative note that (by (24)) $\Delta_n \geq 0 \leftrightarrow T_{n-1} \leq -\frac{d_{n-1}}{n^{1+p}b_n}$. It is again a straightforward calculus argument to show that $-\frac{1}{n^{1+p}b_n} \to \infty$. Since by (25) $T_n \leq 2/p$, for all n, that coupled with Lemma 1 will complete the proof. Consider

$$-j^{1+p}b_j = [1 - \left(\frac{j-1}{j}\right)^p (1 + p/j)]j^{1+p}. \tag{26}$$

Again, let $x = 1/j$ and so $-j^{1+p}b_j$ becomes

$$g(x) = [1 - (1-x)^p(1+px)]/x^{1+p}. \tag{27}$$

We need to show that $g(x) \to 0$ as $x \to 0$. By L'hospital's rule

$$\begin{aligned}
\lim_{x\to 0} g(x) &= \lim_{x\to 0} \frac{-(1-x)^p p + p(1-x)^{p-1}(1+px)}{(1+p)x^p} \\
&= \lim_{x\to 0} \frac{p(1-x)^{p-1}(1+px-1+x)}{(1+p)x^p} \\
&= \lim_{x\to 0} p\frac{(1-x)^{p-1}x(1+p)}{(1+p)x^p} \\
&= \lim_{x\to 0} p(\frac{x}{1-x})^{1-p} = 0. \quad \blacksquare
\end{aligned}$$

We just showed that $E(L_n/n^p)$ converges as $n \to \infty$. Next we show that L_n/n^p has an almost sure limit. We prove this by showing that $L_n/(n+1)^p$ is a (positive) submartingale and that $Var(L_n/(n+1)^p)$ is bounded.

Theorem 3 $\lim_{n\to\infty} E(L_n^2/n^{2p})$ *exists and is finite.*

Proof. Let $U_n = E(L_n^2/n^{2p})$. We first show that there exist constants $0 < c_1(p) < c_2(p) < \infty$ such that for all $n \geq 1$

$$c_1(p) < U_n < c_2(p). \tag{28}$$

The left side of inequality (28) follows trivially from Theorem 2, since $U_n \geq (EL_n/n^p)^2 \to c_p^2$. For the right side inequality of (28), note that

$$\begin{aligned}
E(L_n^2|\mathcal{F}_{n-1}) &= L_{n-1}^2\left(1 - \frac{\lceil pL_{n-1}\rceil}{n}\right) + (L_{n-1}+1)^2\frac{\lceil pL_{n-1}\rceil}{n} \\
&\leq L_{n-1}^2 + (2L_{n-1}+1)\frac{pL_{n-1}+1}{n} \\
&= L_{n-1}^2\left(1 + \frac{2p}{n}\right) + L_{n-1}\frac{p+2}{n} + \frac{1}{n}.
\end{aligned} \tag{29}$$

Thus

$$U_n \le U_{n-1}\left(\frac{n-1}{n}\right)^{2p}\left(1+\frac{2p}{n}\right)+\frac{EL_{n-1}(p+2)}{n^{1+2p}}+\frac{1}{n^{1+2p}}.$$

Therefore,

$$U_n - U_{n-1} < U_{n-1}\left\{\left(\frac{n-1}{n}\right)^{2p}\left(1+\frac{2p}{n}\right)-1\right\}+\frac{E(L_{n-1}/(n-1)^p)(p+2)}{n^{1+p}}+\frac{1}{n^{1+2p}}. \quad (30)$$

Note that from the bound on $f'(x)$ used in the proof of Theorem 2 with $2p$ replacing p $\left(\frac{n-1}{n}\right)^{2p}\left(1+\frac{2p}{n}\right)-1<0$. Since $E(L_{n-1}/(n-1)^p)$ is bounded, it follows from (30) that (with $U_0=0$)

$$U_n=\sum_{j=1}^{n}(U_j-U_{j-1})<\sum_{j=1}^{\infty}\frac{const}{j^{1+p}}<\infty,$$

which accounts for (28).
Now denote $\Delta_j = U_j - U_{j-1}$, so that $U_n=\sum_{j=1}^n \Delta_j$. By virtue of (28) to complete the proof it suffices to show that $\Delta_j>0$ for all j sufficiently large. By (29)

$$E(L_n^2|\mathcal{F}_{n-1})=L_{n-1}^2+2(L_{n-1}+1)\frac{\lceil pL_{n-1}\rceil}{n}\ge L_{n-1}^2+(2L_{n-1}+1)\frac{pL_{n-1}}{n}.$$

Thus

$$\Delta_j \ge U_{j-1}\left\{\left(\frac{j-1}{j}\right)^{2p}\left(1+\frac{2p}{j}\right)-1\right\}+\frac{pE(L_{j-1}/j^p)}{j^{1+p}}.$$

Now for some $0<\theta<1$

$$\left(\frac{j-1}{j}\right)^{2p}=\left(1-\frac{1}{j}\right)^{2p}=1-\frac{2p}{j}+\frac{p(2p-1)}{j^2}\left(1-\frac{\theta}{j}\right)^{-2(1-p)}.$$

Hence there exists a constant $c_3>0$ such that for all $j\ge 1$

$$\left(\frac{j-1}{j}\right)^{2p}\left(1+\frac{2p}{j}\right)-1>-\frac{c_3}{j^2}.$$

Also, there exists a constant $c_4>0$ such that $E(L_{j-1}/j^p)>c_4$ for all $j>1$. But then, for all j sufficiently large, $\Delta_j \ge \frac{-c_3c_2(p)+pc_4j^{1-p}}{j^2}>0$. ∎

Corollary 1 $\lim_{n\to\infty} Var(L_n/n^p)$ *exists and is finite.*

Theorem 4 $\frac{L_n}{(n+1)^p}$ *is a submartingale that converges as* $n\to\infty$ *almost surely to a non-degenerate finite random variable* Λ *such that* $\lim_{n\to\infty}E(L_n/(n+1)^p)=E\Lambda<\infty$, *for all* $0<p\le 1$.

Proof Let $j_n = \lceil pL_n \rceil$.

$$E(L_n|\mathcal{F}_{n-1}) = \frac{j_{n-1}}{n}(L_{n-1}+1) + (1-\frac{j_{n-1}}{n})L_{n-1} = L_{n-1}(1+\frac{p}{n}) + \frac{j_{n-1}-pL_{n-1}}{n}$$

so

$$\begin{aligned} E\left(\tfrac{L_n}{(n+1)^p}|\mathcal{F}_{n-1}\right) = & \frac{L_{n-1}}{n^p}\left(\frac{n}{n+1}\right)^p\left(1+\frac{p}{n}\right) + \frac{j_{n-1}-pL_{n-1}}{n(n+1)^p} \\ & \geq \frac{L_{n-1}}{n^p}\left[\left(\frac{n}{n+1}\right)^p\left(1+\frac{p}{n}\right)\right] \\ & \geq \frac{L_{n-1}}{n^p}. \end{aligned}$$

Therefore, $L_n/(n+1)^p$ is a positive submartingale. Because $E(L_n/(n+1)^p)$ and $E(L_n^2/(n+1)^{2p})$ are both bounded (by virtue of Theorems 2 and 3), Theorem 4 follows from the submartingale convergence theorem.

5 The Quality of the Retained Group of Observations Acquired by a $p-$Percentile Rule

In the previous sections, the focus was on the size of the group retained by the $p-$percentile rule. Here, attention is focused on its quality.

In general $p-$percentile rules yield a qualitative crop. A prime indication of this is Theorem 1 - after n observations of which L_n have been retained, the best $\lceil pL_n \rceil$ of all n observations seen heretofore are among the retained set. As will be shown below in this section, the other retained observations are generally also of high quality.

To this end, the following theorem considers the average rank of the retained items A_n, which equals Q_n/L_n.

Theorem 5 *There exist constants* $0 < b_p < \infty$ *such that for* $0 < p \leq 1$,
$E(A_n)/a_n(p) \rightarrow_{n\rightarrow\infty} b_p$
where

$$a_n(p) = \begin{cases} n^{1-p} & if \quad p < 1/2 \\ n^{1/2}logn & if \quad p = 1/2 \\ n^p & if \quad p > 1/2 \end{cases}$$

and

$$b_p = \begin{cases} c_{1/2}/8 & if \quad p = 1/2 \\ \frac{p^2}{2(2p-1)}c_p & if \quad p > 1/2 \end{cases}, \text{ where } c_p \text{ is the limit of } E(L_n/n^p).$$

Proof: From Assertion iii) at the end of Section 2, with $r(L_n) = j_n = \lceil pL_n \rceil$,

$$E\left(A_{n+1}|\,\mathcal{F}_n\right) = A_n\left[1 + \frac{1 + L_n - j_n}{(n+1)(L_n+1)}\right] + \frac{j_n(j_n-1)/2}{(n+1)L_n}. \tag{31}$$

Let $Y_n = \frac{A_n}{n^{1-p}}$. Equation (31) implies

$$E\left(Y_{n+1}|\mathcal{F}_n\right) = G_n Y_n + B_n \tag{32}$$

where

$$G_n = \left(\frac{n}{n+1}\right)^{1-p}\left(1 + \frac{1 + L_n - j_n}{(n+1)(L_n+1)}\right) \tag{33}$$

and

$$B_n = \frac{j_n(j_n-1)/2}{L_n(n+1)^{2-p}}. \tag{34}$$

We consider B_n first. Since $pL_n \le j_n < pL_n + 1$,

$$\frac{(p^2L_n - p)/2}{(n+1)^{2-p}} \le B_n < \frac{(p^2L_n + p)}{2(n+1)^{2-p}}. \tag{35}$$

By Theorem 2, $EL_n/n^p \overrightarrow{\;n\to\infty\;} c_p$, which implies

$$EB_n n^{2-2p} \overrightarrow{\;n\to\infty\;} p^2c_p/2. \tag{36}$$

We consider G_n next. Let $e_n = pL_n + p - \lceil pL_n \rceil$, so that

$$\frac{1 + L_n - j_n}{L_n + 1} = 1 - p + \frac{e_n}{L_n + 1}. \tag{37}$$

Since $(\frac{n}{n+1})^{1-p} = 1 - \frac{1-p}{n+1} + O(\frac{1}{n^2})$ and since $|e_n| \le 1$,

$$G_n = 1 + \frac{e_n}{(L_n+1)(n+1)} + O\left(\frac{1}{n^2}\right) \tag{38}$$

where $O(\frac{1}{n^2})$ contains elements that multiply e_n, but nevertheless $|n^2O(\frac{1}{n^2})|$ is bounded in n since $|e_n| \le 1$. Substituting equation (38) into equation (32) yields

$$E(Y_{n+1}|\mathcal{F}_n) = Y_n + \left[\frac{e_n}{(L_n+1)(n+1)} + O\left(\frac{1}{n^2}\right)\right]Y_n + B_n. \tag{39}$$

After taking expectations in equation (39), it follows that

$$EY_{n+1} = \sum_{m=0}^{n}\left[EY_{m+1} - EY_m\right] = \sum_{m=1}^{n} ED_m + \sum_{m=1}^{n} EB_m \tag{40}$$

where $D_m = \left[\frac{e_m}{(L_m+1)(m+1)} + O\left(\frac{1}{m^2}\right)\right]Y_m$ and $Y_0 = 0$.

Our aim is to show that $\sum_{m=1}^{n} ED_m$ and $\sum_{m=1}^{n} EB_m$ (or variants thereof for $p \ge \frac{1}{2}$) have finite limits as $n \to \infty$. For the first sum, since $L_m \le m$, it is sufficient to show that $E\sum_{m=0}^{n}\frac{Y_m}{(L_m+1)(m+1)}$ has a finite limit. Now:

$$E\left(\frac{Y_m}{(L_m+1)(m+1)}\right) = E\left(\frac{A_m}{m^{1-p}(m+1)(L_m+1)}\right)$$
$$\leq \frac{1}{m^{2-p}}\left\{E\left(A_m\frac{1}{m^\epsilon}I\{L_m \geq m^\epsilon\}\right) + E\left(A_m I\{L_m < m^\epsilon\}\right)\right\}. \quad (41)$$

By virtue of Lemma A.1 (in the Appendix) there exists a constant $0 < c_{\epsilon,p} < \infty$ such that for $0 < \epsilon < 1/2$

$$P(L_m < m^\epsilon) \leq \frac{c_{\epsilon,p}}{m^{1-\left\lceil 1+\frac{1}{p}\right\rceil\epsilon}} \text{ for all } 1 \leq m < \infty. \quad (42)$$

Note that $A_m \leq m$. Therefore choosing $0 < \epsilon < (1-p)/\gamma_p$ (with $\gamma_p = \left\lceil 1+\frac{1}{p}\right\rceil$), it follows that

$$E\left|\frac{e_m Y_m}{(L_m+1)(m+1)}\right| < \frac{1}{m^{2-p+\epsilon}}E(A_m) + c_{\epsilon,p}/m^{2-p-\gamma_p\epsilon}. \quad (43)$$

We now divide the proof into three cases.

Case(i): $p < 1/2$.

a) $\sum_{m=1}^\infty EB_m < \infty$ by virtue of equation (36).

b) $\sum_{m=1}^\infty \frac{1}{m^{2-p+\epsilon}}E(A_m) = \sum_{m=1}^\infty \frac{1}{m^{1+\epsilon/2}}E\frac{A_m}{m^{1-p+\epsilon/2}} < \infty$ by virtue of Lemma A.2 (in the Appendix).

c) Clearly, $\sum_{m=1}^\infty c_{\epsilon,p}/m^{2-p-\gamma_p\epsilon} < \infty$.

Case(ii): $p = 1/2$.
We need to divide both sides of equation (40) by $\log n$.

a) $\sum_{m=1}^n EB_m/\log n \overrightarrow{n\to\infty} p^2c_p/2 = c_{1/2}/8$ by virtue of equation (36).

b) $\sum_{m=1}^\infty \frac{1}{m^{2-p+\epsilon}}E(A_m) = \sum_{m=1}^\infty \frac{1}{m^{1+\epsilon/2}}E\frac{A_m}{m^{1-p+\epsilon/2}} < \infty$ by virtue of Lemma A.2 (in the Appendix). Hence
$$\frac{\sum_{m=1}^n \frac{1}{m^{2-p+\epsilon}}E(A_n)}{\log n} \underrightarrow{n\to\infty} 0.$$

c) Clearly, $\frac{\sum_{m=1}^n c_{\epsilon,p}/m^{2-p-\gamma_p\epsilon}}{\log n} \underrightarrow{n\to\infty} 0$ (since the numerator is summable). Hence
$$\frac{E(A_n)}{n^{1/2}\log n} \underrightarrow{n\to\infty} c_{1/2}/8.$$

Case(iii): $p > 1/2$.
We need to divide both sides of equation (40) by n^{2p-1}.

a) $\sum_{m=1}^{n} EB_m/n^{2p-1} \overrightarrow{n\to\infty} \frac{p^2}{2(2p-1)}c_p$ by virtue of equation (36).

b) $\frac{\sum_{m=1}^{n} \frac{1}{m^{2-p+\epsilon}}E(A_m)}{n^{2p-1}} = \frac{\sum_{m=1}^{n} \frac{1}{m^{2-2p+\epsilon/2}}E\frac{A_m}{m^{p+\epsilon/2}}}{n^{2p-1}} \overrightarrow{n\to\infty} 0$ by virtue of Lemma A.3 (in the Appendix).

c) For small enough ϵ, $\frac{\sum_{m=1}^{n} c_{\epsilon,p}/m^{2-p-\gamma p\epsilon}}{n^{2p-1}} \overrightarrow{n\to\infty} 0$ (since the numerator is summable). Hence

$$\frac{E(A_n)}{n^p} \overrightarrow{n\to\infty} \frac{p^2}{2(2p-1)}c_p. \quad \blacksquare$$

The theorem below makes use of a result in Robbins and Siegmund(1971) that states: If (for each $n = 1, 2, \cdots$) Y_n, H_n, B_n and C_n are non-negative $\mathcal{F}_n-$ measurable random variables such that

$$E(Y_{n+1}|\mathcal{F}_n) \leq (1 + H_n)Y_n + B_n - C_n$$

where $H_n = (G_n - 1)^+$ in our context, then $\lim_{n\to\infty} Y_n$ exists and is finite and $\sum_{n=1}^{\infty} C_n < \infty$ a.s. on $\left\{\sum_{n=1}^{\infty} H_n < \infty, \sum_{n=1}^{\infty} B_n < \infty\right\}$.

Theorem 6

(i) *If* $0 < p < \frac{1}{2}$, *then* $\frac{A_n}{n^{1-p}}$ *converges almost surely as* $n \to \infty$ *to a non-degenerate random variable.*

(ii) *If* $\frac{1}{2} < p \leq 1$, *then* $Q_n/L_n^2 \overrightarrow{n\to\infty} \frac{p^2}{2(2p-1)}$ *almost surely. Therefore,* $\frac{A_n}{n^p}$ *converges almost surely as* $n \to \infty$ *to a non-degenerate random variable.*

(iii) *If* $p = \frac{1}{2}$, *then* $\frac{Q_n/L_n^2}{\log n} \overrightarrow{n\to\infty} \frac{1}{8}$ *almost surely. Therefore,* $\frac{A_n}{n^{\frac{1}{2}}\log n}$ *converges almost surely as* $n \to \infty$ *to a non-degenerate random variable.*

Proof.

(i) Regard equation (39). Note that B_n of (34) can be written as $B_n = \frac{(p^2L_n+\theta p)/2}{(n+1)^{2-p}}$ where $|\theta| \leq 1$. The almost sure convergence of $\frac{A_n}{n^{1-p}}$ is the result of a direct application of the Theorem of Robbins and Siegmund stated above. The non-degeneracy of the limit follows from the fact that the first observations have rank of order n and their influence on $\frac{A_n}{n^{1-p}}$ does not vanish as $n \to \infty$.

(ii) The idea of the proof is to show that the stochastic process$\{Q_n/L_n^2\}$ gets "pushed downward" when $\frac{Q_n}{L_n^2} > \frac{p^2}{2(2p-1)}$, it gets "pushed upward" when $\frac{Q_n}{L_n^2} < \frac{p^2}{2(2p-1)}$, and there is no real push(upward or downward) when $\frac{Q_n}{L_n^2} \approx \frac{p^2}{2(2p-1)}$.

Denote: $\delta_n = \lceil pL_n \rceil - pL_n = j_n - pL_n$. Note that $0 \leq \delta_n < 1$. Similar to the derivation of (3),

$$
\begin{aligned}
E\left(\frac{Q_{n+1}}{L_{n+1}^2}\middle|\mathcal{F}_n\right) &= \frac{j_n}{n+1}\frac{Q_n+L_n+1}{(L_n+1)^2} \\
&\quad +\frac{n+1-j_n}{(n+1)L_n^2}\left[Q_n+\frac{Q_n-\frac{1}{2}j_n(j_n+1)}{n+1-j_n}-\frac{j_n(L_n-j_n)}{n+1-j_n}\right] \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1}{n+1}-\frac{j_n}{n+1}+\frac{j_n}{n+1}\frac{L_n^2}{(L_n+1)^2}\right] \\
&\quad -\frac{\frac{1}{2}j_n(2L_n-j_n+1)}{(n+1)L_n^2}+\frac{j_n}{(n+1)(L_n+1)} \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1}{n+1}\left(1-\frac{(L_n+1)^2-L_n^2}{(L_n+1)^2}j_n\right)\right] \\
&\quad -\frac{\frac{1}{2}\lceil pL_n\rceil\left(2L_n-\lceil pL_n\rceil+1\right)}{(n+1)L_n^2}+\frac{\lceil pL_n\rceil}{(n+1)(L_n+1)} \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1}{n+1}\left(1-\frac{(2L_n+1)(pL_n+\delta_n)}{(L_n+1)^2}\right)\right] \\
&\quad -\frac{\frac{1}{2}(pL_n+\delta_n)(2L_n-pL_n-\delta_n+1)}{(n+1)L_n^2}+\frac{pL_n+\delta_n+p-p}{(n+1)(L_n+1)} \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1}{n+1}\left(1-\frac{(2(L_n+1)-1)[p(L_n+1)+\delta_n-p]}{(L_n+1)^2}\right)\right] \\
&\quad -\frac{\frac{1}{2}p(2-p)-p}{n+1}-\frac{\frac{1}{2}\delta_n(2-2p)+\frac{1}{2}p}{(n+1)L_n} \\
&\quad +\frac{\frac{1}{2}\delta_n^2-\frac{1}{2}\delta_n}{(n+1)L_n^2}+\frac{\delta_n-p}{(n+1)(L_n+1)} \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1-2p}{n+1}-\frac{1}{n+1}\frac{2\delta_n-3p}{L_n+1}+\frac{\delta_n-p}{(n+1)(L_n+1)^2}\right] \\
&\quad +\frac{\frac{1}{2}p^2}{n+1}-\frac{\delta_n(1-p)+\frac{1}{2}p}{(n+1)L_n}-\frac{\frac{1}{2}\delta_n(1-\delta_n)}{(n+1)L_n^2}+\frac{\delta_n-p}{(n+1)(L_n+1)} \\
&= \frac{Q_n}{L_n^2}\left[1+\frac{1-2p+\frac{\frac{1}{2}p^2}{Q_n/L_n^2}}{n+1}+\frac{3p-2\delta_n+\frac{\delta_n-p}{L_n+1}}{(n+1)(L_n+1)}\right] \\
&\quad +\left[\frac{\delta_n-p}{(n+1)(L_n+1)}-\frac{\delta_n(1-p)+\frac{1}{2}p}{(n+1)L_n}-\frac{\frac{1}{2}\delta_n(1-\delta_n)}{(n+1)L_n^2}\right]. \qquad (44)
\end{aligned}
$$

Note that when L_n is large, the term that can have the greatest influence regarding whether $E\left(\frac{Q_{n+1}}{L_{n+1}^2}\middle|\mathcal{F}_n\right)$ is larger or smaller than $\frac{Q_n}{L_n^2}$ is the term $\frac{1-2p+\frac{\frac{1}{2}p^2}{Q_n/L_n^2}}{n+1}$, and that

$$
1-2p+\frac{\frac{1}{2}p^2}{Q_n/L_n^2} \geq,< 0 \Leftrightarrow \frac{Q_n}{L_n^2} \leq,> \frac{p^2}{2(2p-1)}.
$$

Also note that

1. $Q_{n+1} = Q_n + L_n + 1$ if $L_{n+1} = L_n + 1$
2. $Q_n \le Q_{n+1} \le Q_n + L_n - \lceil pL_n \rceil$ if $L_{n+1} = L_n$.

Therefore,

1. if $L_{n+1} = L_n + 1$, then $\frac{Q_{n+1}}{L_{n+1}^2} = \frac{Q_n}{L_{n+1}^2} + \frac{1}{L_n+1}$,
2. if $L_{n+1} = L_n$, then $0 \le \frac{Q_{n+1}}{L_{n+1}^2} - \frac{Q_n}{L_n^2} \le \frac{1-p}{L_n}$

so that in both cases, if $\frac{Q_n}{L_n^2} \le c_0$ for some $\frac{1}{2} \le \frac{p^2}{2(2p-1)} < c_0 < \infty$, then (since $Q_n \ge L_n(L_n+1)/2$)

$$\begin{aligned} \frac{1}{L_n+1} > \frac{Q_{n+1}}{L_{n+1}^2} - \frac{Q_n}{L_n^2} &\ge \frac{1}{L_n+1} - \frac{2L_n+1}{(L_n+1)^2}\frac{Q_n}{L_n^2} \\ &\ge \frac{1}{L_n+1} - \frac{2L_n+1}{(L_n+1)^2}c_0 \\ &= -\frac{(2c_0-1)L_n + c_0 - 1}{(L_n+1)^2} \\ &> -\frac{2c_0}{L_n}. \end{aligned} \tag{45}$$

It follows that when $\frac{Q_n}{L_n^2} \le c_0$

$$\left| \frac{Q_{n+1}}{L_{n+1}^2} - \frac{Q_n}{L_n^2} \right| < \frac{2c_0}{L_n}. \tag{46}$$

Let $0 < \epsilon < 1$ and define $c_0 = \frac{p^2}{2(2p-1)} + \epsilon$. It follows from equation (44) that when $L_n > \frac{280}{\epsilon}$

$$E\left(\frac{Q_{n+1}}{L_{n+1}^2} \middle| \mathcal{F}_n \right) - \frac{Q_n}{L_n^2} < 0 \quad \text{on} \quad \frac{Q_n}{L_n^2} \ge c_0 \tag{47}$$

and from equation (46) that

$$\left| E\left(\frac{Q_{n+1}}{L_{n+1}^2} \middle| \mathcal{F}_n \right) - \frac{Q_n}{L_n^2} \right| < \frac{2c_0}{L_n} \quad \text{on} \quad \frac{Q_n}{L_n^2} < c_0. \tag{48}$$

Now note that $\frac{Q_1}{L_1^2} = 1$. Let $\zeta_0 = \xi_0 = 0$, and define $\{\zeta_i, \xi_i\}_{i=1}^{\infty}$ recursively as follows:

If $p^2 > 2(2p-1)$ then $\xi_1 = 1$ and for $i \ge 1$

1. $\zeta_i = \min\left\{ n \middle| n > \xi_i, \frac{Q_n}{L_n^2} \ge \frac{p^2}{2(2p-1)} \right\}$; $\zeta_i = \infty$ if $\xi_i = \infty$ or if no such n exists.
2. $\xi_{i+1} = \min\left\{ n \middle| n > \zeta_i, \frac{Q_n}{L_n^2} < \frac{p^2}{2(2p-1)} \right\}$; $\xi_{i+1} = \infty$ if $\zeta_i = \infty$ or if no such n exists.

If $p^2 \le 2(2p-1)$ then $\zeta_1 = 1$ and for $i \ge 1$

1. $\xi_i = \min\left\{n \,\middle|\, n > \zeta_i, \frac{Q_n}{L_n^2} < \frac{p^2}{2(2p-1)}\right\}$; $\xi_i = \infty$ if $\zeta_i = \infty$ or if no such n exists.
2. $\zeta_{i+1} = \min\left\{n \,\middle|\, n > \xi_i, \frac{Q_n}{L_n^2} \geq \frac{p^2}{2(2p-1)}\right\}$; $\zeta_{i+1} = \infty$ if $\xi_i = \infty$ or if no such n exists.

Let $1 \leq j \leq \infty$.
For p such that $p^2 \leq 2(2p-1)$ define

$$Y_n = \begin{cases} \frac{Q_n}{L_n^2} - \sum_{i=1}^{j-1} \frac{2c_0}{(\xi_{i+1})L_{\xi_i}} & \text{for } n = \zeta_j, \zeta_j + 1, \cdots, \xi_j \\ \frac{Q_{\xi_j}}{L_{\xi_j}^2} - \sum_{i=1}^{j} \frac{2c_0}{(\xi_{i+1})L_{\xi_i}} & \text{for } n = \xi_j + 1, \cdots, \zeta_{j+1} - 1. \end{cases}$$

For p such that $p^2 > 2(2p-1)$ define

$$Y_n = \begin{cases} \frac{Q_n}{L_n^2} - \sum_{i=1}^{j} \frac{2c_0}{(\xi_{i+1})L_{\xi_i}} & \text{for } n = \zeta_j, \zeta_j + 1, \cdots, \xi_{j+1} \\ \frac{Q_{\xi_j}}{L_{\xi_j}^2} - \sum_{i=1}^{j-1} \frac{2c_0}{(\xi_{i+1})L_{\xi_i}} & \text{for } n = \xi_j + 1, \cdots, \zeta_j - 1. \end{cases}$$

The series $\{Y_n\}_{n=1}^{\infty}$ satisfies the conditions of Robbins and Siegmund (1971), so it converges a.s. to a finite limit. Since $\sum_{i=1}^{\infty} \frac{1}{(i+1)L_i} < \infty$ almost surely, it follows that $\{\frac{Q_n}{L_n^2} \vee \frac{p^2}{2(2p-1)}\}_{n=1}^{\infty}$ has an a.s. finite limit on the event $\{\zeta_i < \infty \text{ for all } i < \infty\}$, and that $\{\frac{Q_n}{L_n^2}\}_{n=1}^{\infty}$ has an a.s. finite limit on the event $\{\xi_i < \infty, \zeta_i = \infty \text{ for some } i < \infty\}$ when $p^2 > 2(2p-1)$ and on the event $\{\xi_i < \infty, \zeta_{i+1} = \infty \text{ for some } i < \infty\}$ when $p^2 \leq 2(2p-1)$.

For the part of the $\frac{Q_n}{L_n^2}$ that is below $\frac{p^2}{2(2p-1)}$, note that it follows from equation (46) that any upcrossing of $\frac{Q_n}{L_n^2}$ from below $\frac{p^2}{2(2p-1)}$ to above $\frac{p^2}{2(2p-1)}$ will first lead to a value of $\frac{Q_n}{L_n^2}$ that is less than c_0 whenever $L_n > \frac{2c_0}{\epsilon}$.

In a manner similar to the former argument,
for p such that $p^2 \leq 2(2p-1)$ define

$$W_n = \begin{cases} \frac{p^2}{2(2p-1)} - \left(\frac{Q_n}{L_n^2} - \sum_{i=1}^{j} \frac{2c_0}{(\zeta_i+1)L_{\zeta_i}}\right) & \text{for } n = \xi_j, \xi_j + 1, \cdots, \zeta_{j+1} \\ \frac{p^2}{2(2p-1)} - \left(\frac{Q_{\zeta_j}}{L_{\zeta_j}^2} - \sum_{i=1}^{j-1} \frac{2c_0}{(\zeta_i+1)L_{\zeta_i}}\right) & \text{for } n = \zeta_j + 1, \cdots, \xi_j - 1 \end{cases}$$

and for p such that $p^2 > 2(2p-1)$ define

$$W_n = \begin{cases} \frac{p^2}{2(2p-1)} - \left(\frac{Q_n}{L_n^2} - \sum_{i=1}^{j-1} \frac{2c_0}{(\zeta_i+1)L_{\zeta_i}}\right) & \text{for } n = \xi_j, \xi_j + 1, \cdots, \zeta_j \\ \frac{p^2}{2(2p-1)} - \left(\frac{Q_{\zeta_j}}{L_{\zeta_j}^2} - \sum_{i=1}^{j} \frac{2c_0}{(\zeta_i+1)L_{\zeta_i}}\right) & \text{for } n = \zeta_j + 1, \cdots, \xi_j - 1. \end{cases}$$

The series $\{W_n\}$ satisfies the conditions of Robbins and Siegmund (1971), so it converges a.s. to a finite limit. Since $\sum_{i=1}^{\infty} \frac{1}{(i+1)L_i} < \infty$ almost surely, it follows that $\{\frac{Q_n}{L_n^2} \wedge \frac{p^2}{2(2p-1)}\}_{n=1}^{\infty}$ has an a.s. finite limit on the event $\{\xi_i < \infty \text{ for all } i < \infty\}$ and

that $\{\frac{Q_n}{L_n^2}\}_{n=1}^{\infty}$ has an a.s. finite limit on the event $\{\zeta_i < \infty, \xi_i = \infty \text{ for some } i < \infty\}$ when $p^2 > 2(2p-1)$ and on the event $\{\zeta_i < \infty, \xi_{i+1} = \infty \text{ for some } i < \infty\}$ when $p^2 \leq 2(2p-1)$.

Note that $\{\zeta_i < \infty \text{ for all } i < \infty\}$=$\{\xi_i < \infty \text{ for all } i < \infty\}$, and since both $\{Y_n\}$ and $\{W_n\}$ converge a.s. to a finite limit (on this set, too), necessarily $\frac{Q_n}{L_n^2} \overrightarrow{n\to\infty} \frac{p^2}{2(2p-1)}$ a.s. on this set.

Therefore: we have proven that $\frac{Q_n}{L_n^2}$ converges almost surely to a finite limit. It remains to show that his limit is $\frac{p^2}{2(2p-1)}$.

Towards this end, note that by equation (47), once L_n becomes greater than $\frac{280}{\epsilon}$, the variables $\frac{Q_n}{L_n^2}$ constitute a positive supermartingale as long as it exceeds c_0. It is easy to see that once L_n becomes large enough, by virtue of equation (44) there exists a constant $c_1 > 0$ such that when $\frac{Q_n}{L_n^2} \geq c_0$

$$E\left(\frac{Q_{n+1}}{L_{n+1}^2}\middle|\mathcal{F}_n\right) - \frac{Q_n}{L_n^2} \leq -\frac{c_1}{n+1}. \tag{49}$$

Therefore, (since $\sum_{n=m}^{\infty} -\frac{c_1}{n+1} = -\infty$,) it follows that if $\tau_m = \min\{n | n > m, \frac{Q_n}{L_n^2} < c_0\}$, then necessarily $P(\tau_m = \infty | \frac{Q_m}{L_m^2} \geq c_0) = 0$. Since ϵ was arbitrary, it follows that $\overline{\lim}_{n\to\infty} \frac{Q_n}{L_n^2} \leq \frac{p^2}{2(2p-1)}$ almost surely. A similar argument obtains that $\frac{Q_n}{L_n^2}$ is a positive bounded submartingale as long as it is below $\frac{p^2}{2(2p-1)} - \epsilon$ once L_n is large enough, and $E\left(\frac{Q_{n+1}}{L_{n+1}^2}\middle|\mathcal{F}_n\right) \geq \frac{c_2}{n+1}$ for some constant $c_2 > 0$. An argument analogous to the above obtains $\underline{\lim}_{n\to\infty} \frac{Q_n}{L_n^2} \geq \frac{p^2}{2(2p-1)}$ almost surely. Hence we have shown that $\frac{p^2}{2(2p-1)}$ is the a.s. limit of $\frac{Q_n}{L_n^2}$.

(iii) For $p = \frac{1}{2}$, note that

$$\frac{\log n}{\log(n+1)} = \frac{\log n}{\log n + \log(1+1/n)} = 1 - \frac{1+o(1)}{n \log n}. \tag{50}$$

Applying this to equation (44) with $p = \frac{1}{2}$, obtain

$$\begin{aligned} & E\left(\frac{Q_{n+1}/L_{n+1}^2}{\log(n+1)}\middle|\mathcal{F}_n\right) \\ = & \frac{Q_n/L_n^2}{\log n}\left[1 - \frac{1+o(1)}{n\log n} + \frac{1}{8(n+1)\log n}\frac{1}{\frac{Q_n/L_n^2}{\log n}} + \frac{3/2 - 2\delta_n + \frac{\delta_n - \frac{1}{2}}{L_n+1}}{(n+1)(L_n+1)}\right] + \\ & \left[\frac{\delta_n - \frac{1}{2}}{(n+1)(L_n+1)} - \frac{\frac{1}{2}\delta_n + 1/4}{(n+1)L_n} - \frac{\frac{1}{2}\delta_n(1-\delta_n)}{(n+1)L_n^2}\right]. \end{aligned} \tag{51}$$

The situation here is similar to the one in (ii) :

- $E\left(\frac{Q_{n+1}/L_{n+1}^2}{\log(n+1)}\middle|\mathcal{F}_n\right) < \frac{Q_n/L_n^2}{\log n}$ when $\frac{Q_n/L_n^2}{\log n} > \frac{1}{8}+\epsilon$ (and L_n is large enough),
- $E\left(\frac{Q_{n+1}/L_{n+1}^2}{\log(n+1)}\middle|\mathcal{F}_n\right) > \frac{Q_n/L_n^2}{\log n}$ when $\frac{Q_n/L_n^2}{\log n} < \frac{1}{8}-\epsilon$
- $\left|\frac{Q_{n+1}/L_{n+1}^2}{\log(n+1)} - \frac{Q_n/L_n^2}{\log n}\right| < \frac{const}{nL_n\log n}$ when 1/8 lies between $\frac{Q_n/L_n^2}{\log n}$ and $\frac{Q_{n+1}/L_{n+1}^2}{\log(n+1)}$. The proof of (iii) is analogous to that of (ii) (with $\frac{c_1}{[n\log n]}$ replacing $c_1/(n+1)$ in equation (49). The details are omitted. ∎

Remark Our results demonstrate clearly that, when using a $p-$percentile rule, there are two forces at play: the initial observations that are retained and the rest of the retained set. Since L_n is of order n^p, the contribution of the first few observations to the average rank is of order n^{1-p}(since their ranks are typically of order n). Most of the observations will have relatively small rank, and hence contribute n^p to the average. Therefore, when p is small, the first few observations dominate, whereas, when p is large, a sufficient number of observations is retained to dilute the effect of the initial items that are kept on the average rank. (When $p=\frac{1}{2}$, although the contribution of the very first few retained observations vanishes, the first part of the observations that are kept and the other retained items both contribute to the average rank.) The effect is so strong that, when $p>\frac{1}{2}$, the average rank is essentially proportional to the size of the retained set of observations.

6 Simulations

The results in the previous sections describe the performance of rank-based rules; specifically, rules that retain an item if its rank is sufficiently good compared to items already kept. The class of $p-$percentile rules, that essentially retains items in the top $p-$percentile among items already kept, is showcased.

Let n refer to the number of observed items with L_n and $A_n = Q_n/L_n$ denoting the number of items kept and average rank of the retained items, respectively, as above. Note that this average rank is from the retained L_n of the n ranks that are observed to that point. Hence the rank of an item might change as we observe additional items. The results can be summarized as follows:
i) Remarkably, if the observed items are ranked from 1 (best) to n (worst), then *necessarily* the best $\lceil pL_n \rceil$ items will be kept.
ii) As $n \to \infty$, L_n/n^p converges almost surely to a non-degenerate distribution and $E(L_n)/n^p$ converges to a constant.

iii) As $n \to \infty$, $A_n/a_n(p)$ converges almost surely to a non-degenerate distribution and $E(A_n)/a_n(p)$ converges to a constant where $a_n(p) = n^{1-p}$ if $p < 1/2$, $a_n(1/2) = n^{1/2}\log n$ and $a_n(p) = n^p$ if $p > 1/2$.

The aim of this section is to illustrate these results by simulation. To this end, the number of items that are seen is fixed to be $n = 10,000$. For each sample k, the ranks of the items that are kept using 20 different percentiles rules corresponding to $p = .05j$ for $j = 1, \ldots, 20$, are recorded. These results are summarized into $L(p,k)$ and $A(p,k)$ - the number of items retained and the average rank (from among 1 to 10,000) of the retained items using the $p-$percentile rule for replication k. This process was replicated 10,000 times so that k varies from 1 to 10,000.

A representative replication for the median rule is presented in Table 1 to underscore the remarkable property inherent in the first result. Since the average number of items that are kept is approximately 118, the first replication with exactly 118 retained observations is featured.

In Panel 1 of Table 1, where the 118 ranks are kept and sorted, it is evident that
i) The best 59 (half of 118) from among the 10,000 that are seen are kept.
ii) Even if an item is not among the top 59, if it is kept it is still a relatively good item. 83 out of 118 (70%) of the items that are kept are among the top 1% (rank within the first 100 out of 10,000 observations). In addition, only 4 out of 118 (3.4%) of the items that are kept are not in the top 10%.

In Panel 2 of Table 1, the ranks are displayed in the order in which they are kept. Interestingly, all items that are kept after the 4^{th} item is retained are in the top 10% of all items. Since there is a tendency for the items that are retained later to be better than those that are retained towards the beginning, a modification of the median rule that "keeps" for comparative purposes, but does not formally retain the first m items, should perform very well.

Finally, the implication of result 1 that items that are retained are very good ones, is evident from Table 2. The fraction of times the retained rank R_i^n is less than r, for r ranging from $100, 200, \cdots, 10,000$, is computed for all 20 percentile rules across all 10,000 replications. For example, for the median rule, about 70% of the items that are kept have ranks within the first 100 and only 3.4% of the retained ranks are worse than 1,000.

The first column of Table 3 defines the specific percentile rule. The second and third columns illustrate the results for the number of items that are retained as appears in Section 3.

Column 2 reflects Theorem 2. Alongside each value of p is an estimate of M_n/n^p for $n = 10,000$, where M_n is estimated by averaging L_n over the $10,000$ replications. The number immediately below it is the standard deviation of L_n/n^p over the $10,000$ replications. For example, for the median rule ($p = 1/2$), M_n/n^p is estimated to be 1.18. Since the (observed) standard deviation is 0.77, the standard error is $0.77/\sqrt{10,000} = 0.0077$.

Has M_n/n^p essentially converged to its limit at $n = 10,000$? To answer this question, we plotted the estimated M_n/n^p for $n = 1,000, 2,000, \cdots, 10,000$ in Figure 1. Since the lines are essentially horizontal for $p \geq 0.25$, it appears that M_n/n^p is approximately c_p even if n is as small as $1,000$. When p is small, the rate at which M_n/n^p converges is apparently

slower. The corresponding graph for the standard deviation (it is not shown) indicates that the variability of L_n/n^p stabilizes already for $n = 1,000$ for all values of p.

Finally, it is known that $L_n/n^p \to U(0,1)$ when $p = 1$. The average of 0.49974 (approximately 0.5) and standard deviation of 0.28933 (approximately $1/\sqrt{12}$) are consistent with this finding. Likewise, under a limiting $U(0,1)$ distribution, the average proportion of observations whose rank (among the 10,000) is less than $r = 10,000x$ should be approximately $x[1 - \ln x]$, which is consistent with the last line of Table 2.

Column 3 provides estimates for the expected average number of items retained. Obviously, we tend to retain more observations as p increases. For example, for the median rule, the average number of items that are retained are about 118 of the $10,000$ items observed. When $p = 1$, essentially half of the observed items are retained.

We now turn to Columns 4 and 5 which illustrate the quality of the retained items in terms of their average rank.

Column 4 reflects Theorem 5. The values along p are estimates of b_p, provided that the expected values of the average rank, suitably normalized by $a_n(p)$, has essentially converged to their corresponding limiting values when $n = 10,000$. For the median rule, the value of 0.21404 is intended to approximate $c_{1/2}/8 \approx 1.18074/8 = 0.1476$. The discrepancy is evident from Figure 2 where the graph for $p = 0.5$ is clearly decreasing suggesting that the limit has not nearly been reached even when $n = 10,000$.

However, when $p = 0.75$, $b_p = (0.75)^2 c_{0.75} \approx 0.56725(0.90699) = 0.5102$. This value is very close to 0.50989 as appear in Table 3. Figure 3 supports this finding as the graph for $p = 0.75$ is nearly horizontal.

Column 5 provides the average over the replications, of the average rank of the retained items. This should be of order n^{1-p} for $p < 1/2$, n^p for $p > 1/2$ and $\sqrt{n}\log n$ when $p = 1/2$. Table 3 supports the conclusion that the average rank of retained items is minimized for the median rule. The average rank of 197 (with standard error of about 1) is smaller than the corresponding values for all other p reported. Since we automatically retain the first observation whose average rank is $5,000$, about $5,000/118.074 = 42.35$, it follows that 42 of the 197 is reflected in the first item alone. The median rule turns out to be superior based on average rank, as it balances best amortizing the first few retained items by keeping a sufficient number of items, while not keeping too many items some of which would necessarily be inferior.

Column 6 supports the finding that Q_n/L_n^2 converge almost surely to a constant for $p > 1/2$. Note that the standard deviation of Q_n/L_n^2 over the $10,000$ replications is virtually zero (and gets smaller) for large p.

The last column of Table 3 presents the correlation between the number of items that are retained and the sum of the ranks of these items across the $10,000$ replications. As anticipated this correlation is positive and fairly high for $p \geq 1/2$. It is again the behavior of the rank of the first retained observation that affects the correlation, particularly when p is small.

Table 1: Description of a Representative Case for the Median Rule

Number retained	Range	Ranks
59	1-59	1,2,. . . 59
16	60-79	60 61 62 64 65 66 67 68 69 70 71 73 74 75 76 79
8	80-100	80 83 86 87 89 91 92 99
7	101-150	102 106 107 115 131 139 144
11	151-200	154 158 159 164 172 177 184 193 195 196 197
4	210-300	211 222 239 252
6	301-500	304 318 375 397 416 456
3	501-1,000	549 742 999
4	$>$ 1,000	1811 1973 2795 6554

118

Panel 1 − Ranks of representative case

Position	Rank of Item
1-10	6554 2795 1973 1811 999 304 742 211 456 197
11-20	416 549 375 131 397 222 86 195 193 239
21-30	318 49 252 70 184 107 144 96 154 196
31-40	43 75 106 139 172 159 44 177 62 37
41-50	164 24 2 83 158 42 67 40 102 92
51-60	4 30 89 55 115 65 1 79 29 39
61-70	5 28 91 8 75 56 87 69 53 16
71-80	9 38 25 76 32 60 17 27 80 57
81-90	58 7 21 20 48 3 61 22 36 14
91-100	47 73 34 66 71 45 46 59 11 15
101-110	41 68 52 26 64 12 23 50 51 54
111-118	6 33 31 19 18 13 10 35

Panel 2 − Ranks of representative case in order generated

Table 2: Fraction Among the Top r Retained Using a Percentile Rule

$p\backslash r$	100	200	300	500	1,000	2,000	3,000	5,000	10,000
0.05	0.5375	0.6078	0.6490	0.7009	0.7709	0.8408	0.8809	0.9315	1.0000
0.10	0.5584	0.6259	0.6654	0.7149	0.7816	0.8482	0.8864	0.9346	1.0000
0.15	0.6351	0.6977	0.7322	0.7735	0.8275	0.8806	0.9107	0.9484	1.0000
0.20	0.6782	0.7389	0.7712	0.8087	0.8561	0.9009	0.9261	0.9572	1.0000
0.25	0.7212	0.7827	0.8136	0.8479	0.8885	0.9244	0.9438	0.9676	1.0000
0.30	0.7526	0.8201	0.8515	0.8843	0.9198	0.9480	0.9621	0.9786	1.0000
0.35	0.7547	0.8342	0.8691	0.9037	0.9386	0.9634	0.9748	0.9865	1.0000
0.40	0.7606	0.8490	0.8848	0.9185	0.9499	0.9708	0.9800	0.9892	1.0000
0.45	0.7326	0.8479	0.8907	0.9282	0.9600	0.9785	0.9857	0.9925	1.0000
0.50	0.7062	0.8463	0.8954	0.9352	0.9658	0.9821	0.9882	0.9938	1.0000
0.55	0.5236	0.7488	0.8440	0.9165	0.9646	0.9855	0.9918	0.9965	1.0000
0.60	0.4344	0.6700	0.7960	0.9019	0.9651	0.9877	0.9936	0.9974	1.0000
0.65	0.3353	0.5512	0.6947	0.8509	0.9573	0.9883	0.9946	0.9981	1.0000
0.70	0.2377	0.4124	0.5478	0.7361	0.9271	0.9854	0.9944	0.9985	1.0000
0.75	0.1884	0.3309	0.4471	0.6251	0.8673	0.9798	0.9940	0.9987	1.0000
0.80	0.1382	0.2462	0.3385	0.4893	0.7428	0.9499	0.9895	0.9987	1.0000
0.85	0.1027	0.1837	0.2544	0.3747	0.5984	0.8568	0.9637	0.9978	1.0000
0.90	0.0803	0.1434	0.1991	0.2958	0.4828	0.7314	0.8790	0.9905	1.0000
0.95	0.0667	0.1165	0.1605	0.2380	0.3912	0.6093	0.7591	0.9332	1.0000
1.00	0.0575	0.0993	0.1364	0.2016	0.3328	0.5234	0.6622	0.8467	1.0000

Table 3: Number of Items and Average Rank Retained Using a Percentile Rule and their standard deviations

p	L_n/n^p	L_n	A_n/a_n	A_n	Q_n/L_n^2	$\text{Corr}(L_n, Q_n)$
0.05	6.19783	9.823	0.15765	994.704	108.86092	0.49332
	1.80894	2.867	0.10206	643.95496	87.76957	
0.10	4.17790	10.494	0.23835	948.888	102.48531	0.50996
	1.55799	3.913	0.15602	621.12665	88.74309	
0.15	3.44086	13.698	0.30052	754.872	70.60250	0.59061
	1.54264	6.141	0.20210	507.65234	80.11162	
0.20	2.67371	16.870	0.40055	634.829	52.57533	0.64345
	1.34555	8.490	0.27218	431.37625	70.29154	
0.25	2.25283	22.528	0.49799	497.990	33.44803	0.72194
	1.23164	12.316	0.33449	334.48999	55.06502	
0.30	2.11095	33.456	0.57821	364.826	17.08678	0.80747
	1.17824	18.674	0.36928	232.99991	38.89251	
0.35	2.04790	51.441	0.70941	282.421	7.52971	0.87310
	1.11932	28.116	0.38147	151.86591	14.27065	
0.40	1.65326	65.817	0.96748	243.020	5.57129	0.90373
	0.97191	38.692	0.51176	128.54832	13.78241	
0.45	1.50674	95.069	1.33261	211.204	3.48498	0.93244
	0.91630	57.815	0.68988	109.33862	12.41476	
0.50	1.18074	118.074	0.21404	197.138	0.31669	0.94194
	0.77037	77.037	0.11236	103.48739	1.34502	
0.55	1.39694	221.400	1.42291	225.516	1.08661	0.95778
	0.77247	122.428	0.70327	44.37333	0.52466	
0.60	1.19805	300.937	0.97761	245.565	0.86203	0.95778
	0.68568	172.235	0.51435	20.47664	0.38195	
0.65	1.09353	435.342	0.74808	297.816	0.70456	0.95740
	0.63277	251.910	0.41435	10.40800	0.15384	
0.70	1.04488	659.275	0.63442	400.292	0.61184	0.95942
	0.57848	364.996	0.34542	5.47454	0.03791	
0.75	0.90699	906.990	0.50989	509.890	0.56518	0.95924
	0.51406	514.060	0.28633	2.86330	0.02036	
0.80	0.84126	1333.307	0.44875	711.221	0.53453	0.96167
	0.46911	743.489	0.24937	1.57342	0.00741	
0.85	0.77886	1956.408	0.40202	1009.829	0.51667	0.96357
	0.42860	1076.595	0.22098	0.87974	0.00459	
0.90	0.69251	2756.931	0.35068	1396.082	0.50677	0.96553
	0.38479	1531.876	0.19476	0.48922	0.00403	
0.95	0.60885	3841.583	0.30455	1921.580	0.49915	0.96760
	0.34306	2164.562	0.17188	0.27241	0.00402	
1.00	0.49974	4997.400	0.24950	2495.000	0.49838	0.96827
	0.28933	2893.300	0.14461	0.14461	0.00364	

- $n = 10,000$ and Number of Replications $= 10,000$.
- $a_n(p) = n^{1-p}$ if $p < .5$; $a_n(p) = n^{1/2}ln(n)$ if $p = 0.5$ and $a_n(p) = n^p$ if $p > .5$.
- First row along p is the average; second row is the std. dev.

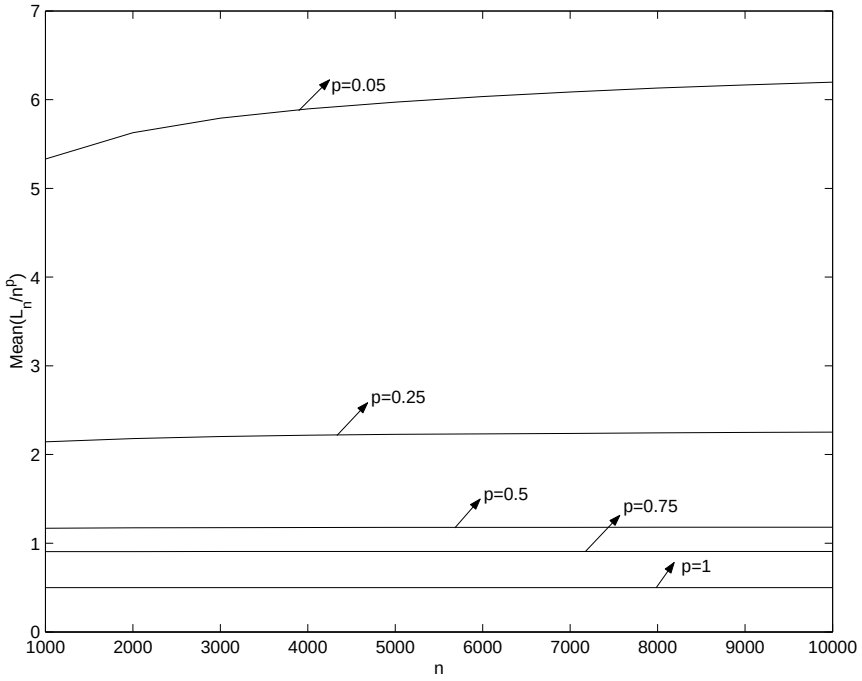

Figure 1: Behavior of normalized average number of items retained

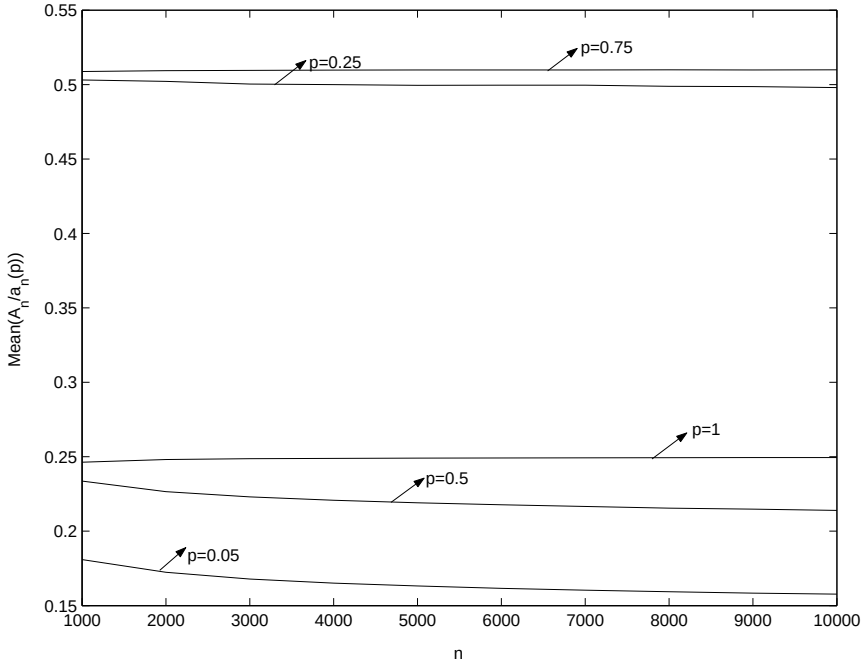

Figure 2: Behavior of normalized average rank

7 Remarks

Remark 1 Some of the results carry over to the inverse problem. Suppose that a $p-$percentile rule is applied. Let Z_i (for $i \geq 1$) be the number of observations made from the instant that the size of the set of retained observations became $i-1$ until its size became i. Also, let $T_n = \sum_{i=1}^{n} Z_i$ be the number of observations made until n have been retained. The results stated in Theorems 4 and 6 carry over directly to T_n and $Q_{T_n} : L_{T_n}/T_n^p = n/T_n^p$ converges a.s. as $n \to \infty$ to a finite, non-degenerate random variable and for $p < \frac{1}{2}, p = \frac{1}{2}, p > \frac{1}{2}$ the quantities

(α) $\frac{Q_{T_n}}{n}/T_n^p,\ \frac{Q_{T_n}}{n}/(T_n^{\frac{1}{2}} \log T_n),\ \frac{Q_{T_n}}{n}/T^{1-p}$

respectively, converge a.s. as $n \to \infty$ to non-degenerate random variables, and for $p > \frac{1}{2}$

(β) $Q_{T_n}/n^{2p} \overrightarrow{\ n\to\infty\ } p^2/[2(2p-1)]$ a.s..

Evaluating expectations of $T_n = \sum_{i=q}^{n} Z_i$ is a more complicated matter.

For $0 < p \leq 1, n \geq 2$ the expectation of T_n is ∞ (since $EZ_2 = \infty$).

For $0 < p \leq 1$, the expectation of each of the three expressions in (α) converges to a finite positive constant as $n \to \infty$, and the expectation of the term on the left side of (β) converges to the value in the right side as $n \to \infty$ for $p > \frac{1}{2}$.

These statements can be proved using methods similar to those used for proving Theorems 3 and 5.

Remark 2 Other rules can be evaluated in a manner similar to the $p-$percentile rules. For example, for $k-$record rules using the Assertion of Section 2, it can be shown that $L_n/\log n$ converges almost surely to k as $n \to \infty$ and $EL_n/\log n$ also converges to k. It can be shown that $Q_n/(n+1)$ is a (non-negative) submartingale that converges a.s. as $n \to \infty$ to a non-degenerate random variable, and $EQ_n/(n+1) \overrightarrow{\ n\to\infty\ } k$. Thus $A_n \log n/n$ converges a.s. to a non-degenerate random variable.

Remark 3 The complexity of the sorting problem is of order $n \log n$. Sometimes, one is interested in retaining (in sorted form) only some of the best observations rather than the whole set. In this case, a $p-$percentile rule with $0 < p < 1$ obtains a sorted set of best observations, and the complexity is of order n. To see this, note that initially each observation has to be compared only to the $p-$percentile of the retained set — amounting to n operations — and each retained observation must be compared to (roughly) $100p\%$ of the retained observations, amounting (at most) to another $2\sum_{j=1}^{L_n} \log(j+1) = O(L_n \log L_n) = O_p(n^p \log n)$ operations (since the retained set can be stored in sorted condition).

A Appendix

Lemma A.1 *For any $p-$percentile rule with $0 < p \leq 1$ and any $0 < \epsilon < 1$, there exists a constant $0 < c_{\epsilon,p} < \infty$ such that*

$$P(L_n < k) \leq c_{\epsilon,p} k^{r_0(1-\epsilon)}/n^{1-\epsilon} \tag{52}$$

where $r_0 = \lceil 1/p \rceil$.

Proof: After $m-1$ observations have been retained by the $p-$percentile rule, let Z_m denote the number of additional observations until the next retention. Note that $Z_1 = 1$. Let $N_n = \sum_{i=1}^{n} Z_i$ be the number of observations made until n items have been retained. Thus,

$$P(L_n < k) = P(N_k > n) = P(N_k^{1-\epsilon} > n^{1-\epsilon}) \leq EN_k^{1-\epsilon}/n^{1-\epsilon}. \tag{53}$$

Without loss of generality, assume that the X_i have a $U[0,1]$ distribution. Let X_i^n denote the observation with rank i among $X_1, X_2, \cdots, X_n$. Note that conditional on $X_1, \cdots, X_{N_m}$, the distribution of Z_{m+1} is

$$\text{Geometric } p \text{ with } p = 1 - X_{j_{N_m}}^{N_m} = 1 - X_{\lceil pm \rceil}^{N_m}.$$

Also note that conditional on N_m, the distribution of $1 - X_{\lceil pm \rceil}^{N_m}$ is Beta $(\lceil pm \rceil, N_m + 1 - \lceil pm \rceil)$.

Therefore, for $0 \leq \epsilon < 1$

$$\begin{aligned}
E\left(Z_{m+1}^{1-\epsilon} \middle| N_m\right) &= E\left[E\left(Z_{m+1}^{1-\epsilon} \middle| N_m, X_{\lceil pm \rceil}^{N_m}\right) \middle| N_m\right] \\
&\leq E\left(\left[E\left(Z_{m+1} \middle| N_m, X_{\lceil pm \rceil}^{N_m}\right)\right]^{1-\epsilon} \middle| N_m\right) \\
&= E\left[\left(\frac{1}{1 - X_{\lceil pm \rceil}^{N_m}}\right)^{1-\epsilon} \middle| N_m\right] \\
&= \frac{N_m!}{(\lceil pm \rceil - 1)!(N_m - \lceil pm \rceil)!} \cdot \frac{\Gamma(\lceil pm \rceil + \epsilon - 1)\Gamma(N_m + 1 - \lceil pm \rceil)}{\Gamma(N_m + \epsilon)} \\
&= \frac{N_m!}{(\lceil pm \rceil - 1)!} \cdot \frac{\Gamma(\lceil pm \rceil + \epsilon - 1)}{\Gamma(N_m + \epsilon)} \\
&\leq \frac{N_m}{\lceil pm \rceil + \epsilon - 1}.
\end{aligned} \tag{54}$$

For $\epsilon = 0$ obtain $E(Z_{m+1}|N_m) \leq N_m/(\lceil pm \rceil - 1)$, so that

$$E(N_{m+1}|N_m) \leq \frac{\lceil pm \rceil}{\lceil pm \rceil - 1} N_m. \tag{55}$$

Hence

$$E(N_{m+1}^{1-\epsilon}|N_m) \leq [E(N_{m+1}|N_m)]^{1-\epsilon} \leq \left(\frac{\lceil pm \rceil}{\lceil pm \rceil - 1}\right)^{1-\epsilon} N_m^{1-\epsilon}. \tag{56}$$

Letting m_p be the smallest m such that $\lceil pm \rceil > 1$, it follows that

$$E(N_{m+1}^{1-\epsilon}|N_{m_p}) \leq N_{m_p}^{1-\epsilon} \prod_{i=m_p}^{m} \left(\frac{\lceil pi \rceil}{\lceil pi \rceil - 1} \right)^{1-\epsilon}. \tag{57}$$

We first show that $EN_{m_p}^{1-\epsilon}$ is finite. By virtue of equation (54),

$$\begin{aligned} E(N_{m+1}^{1-\epsilon}|N_m) &= E((N_m + Z_{m+1})^{1-\epsilon}|N_m) \\ &\leq E(N_m^{1-\epsilon}|N_m) + E(Z_{m+1}^{1-\epsilon}|N_m) \\ &\leq N_m^{1-\epsilon} + E(Z_{m+1}^{(1-\epsilon/2)^2}|N_m) \\ &\leq N_m^{1-\epsilon} + [E(Z_{m+1}^{1-\epsilon/2}|N_m)]^{1-\epsilon/2} \\ &\leq N_m^{1-\epsilon} + \frac{N_m^{1-\epsilon/2}}{(\lceil pm \rceil + \epsilon/2 - 1)^{1-\epsilon/2}}. \end{aligned} \tag{58}$$

This recursive relation can be applied repeatedly. Since $EN_1^{1-\epsilon} = 1$, it follows that $EN_{m_p}^{1-\epsilon}$ is finite for all $\epsilon > 0$.

Finally, note that $\frac{\lceil pi \rceil}{\lceil pi \rceil - 1} > 1$ and can appear (in the product) $\prod_{i=m_p}^{m} \frac{\lceil pi \rceil}{\lceil pi \rceil - 1}$ at most $r_0 = \lceil 1/p \rceil$ times. Hence,

$$\begin{aligned} \prod_{i=m_p}^{m} \left(\frac{\lceil pi \rceil}{\lceil pi \rceil - 1} \right)^{1-\epsilon} &\leq \left(\frac{\lceil pm_p \rceil}{\lceil pm_p \rceil - 1} \cdot \frac{\lceil pm_p + 1 \rceil}{\lceil pm_p \rceil - 1} \cdot \frac{\lceil pm_p + 2 \rceil}{\lceil pm_p \rceil - 1} \cdots \frac{\lceil pm \rceil}{\lceil pm_p \rceil - 1} \right)^{1-\epsilon} \\ &\leq \left(\frac{\lceil pm \rceil}{\lceil pm_p \rceil - 1} \right)^{r_0(1-\epsilon)} = \lceil pm \rceil^{r_0(1-\epsilon)} \end{aligned} \tag{59}$$

For $m = k$, the conclusion follows from equation (53) and (57). ∎

Lemma A.2 *Let $0 < \epsilon < 1$ and consider a $p-$percentile rule with $0 < p \leq \frac{1}{2}$. Then*

$$\lim_{n \to \infty} E\left(\frac{A_n}{n^{1-p+\epsilon}} \right) = 0.$$

Proof. Let $1 \leq k_\epsilon < \infty$ be an integer such that

$$\frac{1 + L_n - j_n}{1 + L_n} < (1-p)(1+\epsilon) \text{ whenever } L_n \geq k_\epsilon.$$

From Lemma A.1 obtain

$$P(L_n < k_\epsilon) \leq c^*_{\epsilon,p,k_\epsilon} / n^{1-\epsilon} \tag{60}$$

Note that $A_n \leq n$ and $j_n \leq L_n$. Therefore

$$\begin{aligned} E\left(A_{n+1}|\,\mathcal{F}_n\right) &= A_n \left(1 + \frac{1 + L_n - j_n}{(n+1)(1+L_n)} \right) + \frac{j_n(j_n - 1)/2}{(n+1)L_n} \\ &\leq A_n + A_n \frac{1 + L_n - j_n}{(n+1)(1+L_n)} [I\{L_n \geq k_\epsilon\} + I\{L_n < k_\epsilon\}] + \frac{p^2 L_n + p}{2(n+1)}. \end{aligned} \tag{61}$$

Now for $0 < p \leq 1/2$, the right hand side of (62) is less than or equal to

$$A_n \left[1 + \frac{(1-p)(1+\epsilon)}{n}\right] + \frac{n}{n+1} I\{L_n < k_\epsilon\} + \frac{p^2 L_n + p}{2(n+1)}.$$

Hence, with $\epsilon < \{p \wedge c_p\}$ and large enough n,

$$\begin{aligned} E\left(A_{n+1}\right) &\leq E\left(A_n\right)\left[1 + \frac{(1-p)(1+\epsilon)}{n}\right] + \frac{c^*_{\epsilon,p,k_\epsilon}}{n^{1-\epsilon}} + \frac{p^2(\epsilon + c_p)}{2n^{1-p}} \\ &\leq E\left(A_n\right)\left[1 + \frac{(1-p)(1+\epsilon)}{n}\right] + \frac{c_p}{n^{1-p}}. \end{aligned} \tag{62}$$

Let $\gamma_1 = 1$ and define $\gamma_{n+1} = \gamma_n / [1 + \frac{(1-p)(1+\epsilon)}{n}]$. Thus, $\{\gamma_n\}$ is a decreasing sequence, and there exists $0 < \gamma_\infty < \infty$ such that

$$\lim_{n\to\infty} n^{(1-p)(1+\epsilon)} \gamma_n = \gamma_\infty.$$

Note that for large enough n

$$\gamma_{n+1} E(A_{n+1}) \leq \gamma_n E(A_n) + \frac{2c_p \gamma_\infty}{n^{2(1-p)+(1-p)\epsilon}}\ . \tag{63}$$

Because $2(1-p) \geq 1$, it follows that

$$\overline{\lim_{n\to\infty}}\ \gamma_n E(A_n) < \infty;\ \textit{i.e.}\ \overline{\lim_{n\to\infty}}\ E\left(\frac{A_n}{n^{1-p+(1-p)\epsilon}}\right) < \infty.$$

Hence

$$\lim_{n\to\infty} E\left(\frac{A_n}{n^{1-p+\epsilon}}\right) = 0. \quad \blacksquare \tag{64}$$

Lemma A.3 *Let $0 < \epsilon$ and consider a $p-$percentile rule with $\frac{1}{2} < p \leq 1$. Then for all $L_n > k_\epsilon$*

$$\lim_{n\to\infty} E\left(\frac{A_n}{n^{p+\epsilon}}\right) = 0.$$

Proof. Let k_ϵ be an integer such that whenever $L_n > k_\epsilon$

$$\frac{1 + L_n - j_n}{1 + L_n} < (1-p)(1+\epsilon) < p(1+\epsilon). \tag{65}$$

By (62) and (66),

$$E\left(A_{n+1}|\,\mathcal{F}_n\right) \leq A_n\left[1 + \frac{p(1+\epsilon)}{n}\right] + \frac{n}{n+1} I\{L_n < k_\epsilon\} + \frac{p^2 L_n + p}{2(n+1)}. \tag{66}$$

Hence, with $\epsilon < p \wedge c_p$ and large enough n,

$$\begin{aligned} E\left(A_{n+1}\right) &\leq E\left(A_n\right)\left[1 + \frac{p(1+\epsilon)}{n}\right] + \frac{c^*_{\epsilon,p,k_\epsilon}}{n^{1-\epsilon}} + \frac{p^2(\epsilon + c_p)n^p}{2n} \\ &\leq E\left(A_n\right)\left[1 + \frac{p(1+\epsilon)}{n}\right] + \frac{c_p}{n^{1-p}}. \end{aligned} \tag{67}$$

Let $\gamma_1 = 1$ and define

$$\gamma_{n+1} = \gamma_n \Big/ \left[1 + \frac{p(1+\epsilon)}{n}\right].$$

Note that $\{\gamma_n\}$ is a decreasing sequence and there exists a constant $0 < \gamma_\infty < \infty$ such that $\lim_{n\to\infty} n^{p(1+\epsilon)}\gamma_n = \gamma_\infty$. Note that

$$\gamma_{n+1} E\left(A_{n+1}\right) \leq \gamma_n E\left(A_n\right) + \gamma_{n+1}\frac{c_p}{n^{1-p}}\,. \tag{68}$$

Therefore, there exists an integer $1 \leq n_\epsilon < \infty$ such that for all $n \geq n_\epsilon$

$$\gamma_{n+1} E\left(A_{n+1}\right) \leq \gamma_n E\left(A_n\right) + \gamma_\infty \frac{2c_p}{n^{1+p\epsilon}}\,. \tag{69}$$

It follows that

$$\overline{\lim_{n\to\infty}}\, \gamma_n E(A_n) < \infty;\ \textit{i.e.}\ \overline{\lim_{n\to\infty}}\, E\left(\frac{A_n}{n^{p+p\epsilon}}\right) < \infty$$

Hence

$$\lim_{n\to\infty} E\left(\frac{A_n}{n^{p+\epsilon}}\right) = 0. \quad \blacksquare \tag{70}$$

References

[1] Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N. (1998) *Records.* Wiley, New York.

[2] Gilbert, J.P. and Mosteller F.(1996). Recognizing the maximum of a sequence. *Journal of the American Statistical Association*, 61, 35-73.

[3] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Sequences and Processes*. Springer-Verlag, Inc. New York.

[4] Preater, J. (1994). A multiple stopping problem. *Probablity in the Engineering and Information Sciences*, 8, 169-177.

[5] Preater, J. (2000). Sequential selection with a better than average rule. *Statistics and Probability Letters*, 50, 187-191.

[6] Robbins, H. and Siegmund, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing Methods in Statistics,* J.S. Rustagi Editor. Academic Press, New York.

[7] Resnick, S.I.(1987). *Extreme Values, Regular Variation and Point Processes.* Springer-Verlag, Inc. New York.