On the variance of sample size

Yuri Yakubovich[1]

Joint work with L. Bogachev and A. Gnedin

3d Northern Triangular Seminar
Euler International Mathematical Institute, St.Petersburg
12 April, 2011

[1] St. Petersburg State University and St. Petersburg University of Film and Television, Russia

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with discrete distribution. We consider the set $\mathcal{S}_n$ of the first n samples and let the r.v. $K_n = |\mathcal{S}_n|$ be its size.
K_n is the number of distinct values among the first n samples.

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with discrete distribution. We consider the set $\mathcal{S}_n$ of the first n samples and let the r.v. $K_n = |\mathcal{S}_n|$ be its size.
K_n is the number of distinct values among the first n samples.
Since values of X_j's are of no importance for us, without loss of generality we may arrange them so that $\mathbb{P}[X_j = x_i] = p_i$ and

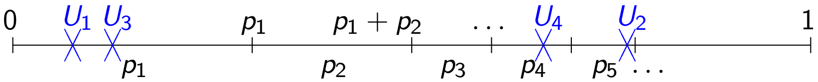
$$p_1 \geq p_2 \geq \cdots > 0, \qquad \sum_i p_i = 1.$$

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with discrete distribution. We consider the set $\mathcal{S}_n$ of the first n samples and let the r.v. $K_n = |\mathcal{S}_n|$ be its size.
K_n is the number of distinct values among the first n samples.
Since values of X_j's are of no importance for us, without loss of generality we may arrange them so that $\mathbb{P}[X_j = x_i] = p_i$ and

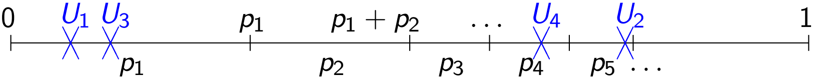
$$p_1 \geq p_2 \geq \cdots > 0, \qquad \sum\nolimits_i p_i = 1.$$

One can also admit that the underlying distribution has a continuous part, but all samples from continuous distribution are different a.s. and it is simple to analyze. So we always suppose that the distribution of X_j's is purely discrete and its support is infinite. In this case $K_n \to \infty$ as $n \to \infty$, so it is possible (and interesting) to investigate its behaviour in the limit.

It is convenient to model this construction by a sequence of the i.i.d. random variables uniformly distributed on the unit interval $[0, 1]$ divided into subintervals of lengths $p_1, p_2, \dots$:



It is convenient to model this construction by a sequence of the i.i.d. random variables uniformly distributed on the unit interval $[0, 1]$ divided into subintervals of lengths $p_1, p_2, \dots$:

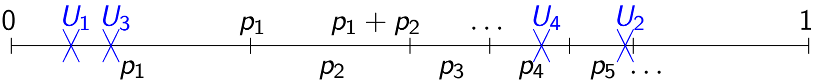


$M_{1,4} = 2$ $\quad$ $M_{2,4} = M_{3,4} = 0$ $\quad$ $M_{4,4} = M_{5,4} = 1$ $\quad$ $M_{6,4} = \cdots = 0$

$I_{1,4} = 1$ $\quad$ $I_{2,4} = I_{3,4} = 0$ $\quad$ $I_{4,4} = I_{5,4} = 1$ $\quad$ $I_{6,4} = \cdots = 0$

- $M_{i,n}$—the number of samples that get into i-th subinterval among the first n samples;
- $I_{i,n} = \mathbb{1}_{M_{i,n}>0}$—the indicator of the event "i-th subinterval is hit by at least one sample among the first n samples".

It is convenient to model this construction by a sequence of the i.i.d. random variables uniformly distributed on the unit interval $[0, 1]$ divided into subintervals of lengths $p_1, p_2, \dots$:



$M_{1,4} = 2$ $\quad$ $M_{2,4} = M_{3,4} = 0$ $\quad$ $M_{4,4} = M_{5,4} = 1$ $\quad$ $M_{6,4} = \cdots = 0$

$I_{1,4} = 1$ $\quad$ $I_{2,4} = I_{3,4} = 0$ $\quad$ $I_{4,4} = I_{5,4} = 1$ $\quad$ $I_{6,4} = \cdots = 0$

- $M_{i,n}$—the number of samples that get into i-th subinterval among the first n samples;
- $I_{i,n} = \mathbb{1}_{M_{i,n}>0}$—the indicator of the event "i-th subinterval is hit by at least one sample among the first n samples".

Obviously,

$$\sum_i M_{i,n} = n, \qquad \sum_i I_{i,n} = K_n.$$

Poissonization is a standard procedure in such problems. It can be described in several alternative but equivalent ways:

- Consider not a fixed number of uniform sample points $U_1, \dots, U_n$ but a random number $U_1, \dots, U_{N(n)}$ where $N(n)$ has the Poisson distribution with parameter n;

Poissonization is a standard procedure in such problems. It can be described in several alternative but equivalent ways:

- Consider not a fixed number of uniform sample points $U_1, \ldots, U_n$ but a random number $U_1, \ldots, U_{N(n)}$ where $N(n)$ has the Poisson distribution with parameter n;
- Consider the system in continuous time and add uniform samples $U_1, U_2, \ldots$ with random independent exponentially distributed delays with mean 1, and stop at the time n;

Poissonization is a standard procedure in such problems. It can be described in several alternative but equivalent ways:

- Consider not a fixed number of uniform sample points $U_1, \dots, U_n$ but a random number $U_1, \dots, U_{N(n)}$ where $N(n)$ has the Poisson distribution with parameter n;
- Consider the system in continuous time and add uniform samples $U_1, U_2, \dots$ with random independent exponentially distributed delays with mean 1, and stop at the time n;
- Consider the points of the Poisson point process (PPP) of intensity n on $[0, 1]$ instead of n uniform samples.

Poissonization is a standard procedure in such problems. It can be described in several alternative but equivalent ways:

- Consider not a fixed number of uniform sample points $U_1, \ldots, U_n$ but a random number $U_1, \ldots, U_{N(n)}$ where $N(n)$ has the Poisson distribution with parameter n;
- Consider the system in continuous time and add uniform samples $U_1, U_2, \ldots$ with random independent exponentially distributed delays with mean 1, and stop at the time n;
- Consider the points of the Poisson point process (PPP) of intensity n on $[0, 1]$ instead of n uniform samples.

We use indices for the fixed n version and brackets notation for the Poissonized version (K_n vs $K(n)$ etc).

The Poissonized version has many advantages:

- The PPP representation shows that for each $n > 0$, $M_i(n)$ form the sequence of independent r.v.'s having the Poisson distribution with mean np_i.

The Poissonized version has many advantages:

- The PPP representation shows that for each $n > 0$, $M_i(n)$ form the sequence of independent r.v.'s having the Poisson distribution with mean np_i.
- The indicators $I_i(n)$ are independent Bernoulli r.v.'s with success probability $1 - e^{-np_i}$.

The Poissonized version has many advantages:

- The PPP representation shows that for each $n > 0$, $M_i(n)$ form the sequence of independent r.v.'s having the Poisson distribution with mean np_i.
- The indicators $I_i(n)$ are independent Bernoulli r.v.'s with success probability $1 - e^{-np_i}$.
- No need for the normalization $\sum_i p_i = 1$: by a linear time change one can renormalize this sum, so just its finiteness is needed.

The Poissonized version has many advantages:

- The PPP representation shows that for each $n > 0$, $M_i(n)$ form the sequence of independent r.v.'s having the Poisson distribution with mean np_i.
- The indicators $I_i(n)$ are independent Bernoulli r.v.'s with success probability $1 - e^{-np_i}$.
- No need for the normalization $\sum_i p_i = 1$: by a linear time change one can renormalize this sum, so just its finiteness is needed.
- Additive structure:
  - (p_i') and (p_i'')—two sequences with finite sums;
  - $(p_i) = (p_i') \cup (p_i'')$—union as multisets;
  - $K'(n)$, $K''(n)$ and $K(n)$—the corresponding numbers of different samples in the Poissonized settings

  then

$$K(n) \stackrel{d}{=} K'(n) + K''(n), \qquad K'(n),\ K''(n) \text{ independent.}$$

We are interested in the mean and variance of the number of different values in the first n samples. Let us introduce

$$\Phi_n = \mathbb{E}[K_n], \qquad V_n = \mathrm{Var}[K_n]$$

and the Poissonized analogs

$$\Phi(n) = \mathbb{E}[K(n)], \qquad V(n) = \mathrm{Var}[K(n)]$$

We are interested in the mean and variance of the number of different values in the first n samples. Let us introduce

$$\Phi_n = \mathbb{E}[K_n], \qquad V_n = \mathrm{Var}[K_n]$$

and the Poissonized analogs

$$\Phi(n) = \mathbb{E}[K(n)], \qquad V(n) = \mathrm{Var}[K(n)]$$

The formulas become particularly simple after Poissonization:

$$\Phi(n) := \mathbb{E}[K(n)] = \sum\nolimits_i (1 - e^{-np_i}),$$
$$V(n) := \mathrm{Var}\, K(n) = \sum\nolimits_i (e^{-np_i} - e^{-2np_i}) = \Phi(n) - \Phi(2n).$$

We shall not use formulas for Φ_n and V_n but one can write them down.

Since there are infinitely many possible values, $K_n \to \infty$ a.s. as $n \to \infty$, and so does its mean: $\Phi_n \to \infty$. It is also known that $K_n/\mathbb{E}[K_n] \to 1$ as $n \to \infty$ in probability (Bahadur, 1960) and even a.s. (Karlin, 1967).

Since there are infinitely many possible values, $K_n \to \infty$ a.s. as $n \to \infty$, and so does its mean: $\Phi_n \to \infty$. It is also known that $K_n/\mathbb{E}[K_n] \to 1$ as $n \to \infty$ in probability (Bahadur, 1960) and even a.s. (Karlin, 1967).
The next question is whether the variance V_n increase to infinity or not. This question is particularly interesting because it is known that if $V_n \to \infty$, $n \to \infty$, then $\frac{K_n - \Phi_n}{\sqrt{V_n}} \Rightarrow \mathcal{N}$, the standard normal distribution (Karlin, 1967; Dutko, 1984; LLT by Hwang and Janson, 2006).

Since there are infinitely many possible values, $K_n \to \infty$ a.s. as $n \to \infty$, and so does its mean: $\Phi_n \to \infty$. It is also known that $K_n/\mathbb{E}[K_n] \to 1$ as $n \to \infty$ in probability (Bahadur, 1960) and even a.s. (Karlin, 1967).
The next question is whether the variance V_n increase to infinity or not. This question is particularly interesting because it is known that if $V_n \to \infty$, $n \to \infty$, then $\frac{K_n - \Phi_n}{\sqrt{V_n}} \Rightarrow \mathcal{N}$, the standard normal distribution (Karlin, 1967; Dutko, 1984; LLT by Hwang and Janson, 2006).
This is easy to see for the Poissonized version of the problem. Then $K(n)$ is a sum of independent r.v.'s with Bernoulli distributions and the asymptotic normality follows, say, by application of Lindeberg's theorem. De-Poissonization requires some work which was done by Dutko in 1984.

So the interesting case is when V_n does not tend to ∞. Two alternatives are possible: either V_n oscillates unboundedly or it remains bounded as $n \to \infty$. Introduce

$$\overline{v} := \limsup_{n\to\infty} V_n, \qquad \underline{v} := \liminf_{n\to\infty} V_n.$$

So the interesting case is when V_n does not tend to ∞. Two alternatives are possible: either V_n oscillates unboundedly or it remains bounded as $n \to \infty$. Introduce

$$\overline{v} := \limsup_{n\to\infty} V_n, \qquad \underline{v} := \liminf_{n\to\infty} V_n.$$

We propose the following criterion for boundedness of V_n:

Theorem 1. The boundedness of $V(t)$ is equivalent to the existence of an integer k such that

$$\limsup_j \frac{p_{j+k}}{p_j} \leq \frac{1}{2}.$$

Moreover, this inequality implies $\overline{v} \leq k$. If for any k

$$\liminf_j \frac{p_{j+k}}{p_j} \geq \frac{1}{2}$$

then $\underline{v} = \infty$.

If V_n remains bounded, it is interesting whether it converges to a limit as $n \to \infty$. It can be also checked in terms of "lagged ratio":

Theorem 2. The limit $\lim_n V_n = v$ exists if and only if

$$\lim_j \frac{p_{j+k}}{p_j} = \frac{1}{2},$$

and in this case $v = k$.

This has an unexpected corollary:

If V_n converges to a finite limit, this limit is a positive integer.

If V_n remains bounded, it is interesting whether it converges to a limit as $n \to \infty$. It can be also checked in terms of "lagged ratio":

Theorem 2. The limit $\lim_n V_n = v$ exists if and only if

$$\lim_j \frac{p_{j+k}}{p_j} = \frac{1}{2},$$

and in this case $v = k$.

This has an unexpected corollary:

If V_n converges to a finite limit, this limit is a positive integer.

This can be extended by considering the case when the distribution has just finitely many atoms. In this (not very interesting) case $K_n \to \mathrm{const}$ a.s. so its variance converges to zero.

Let p_j form the geometric sequence with the common ratio $1/2$, that is $p_j = 1/2^j$.
Then the Poissonized variance can be calculated as follows:

$$V(t) = \lim_{m\to\infty} \sum_{j=1}^{m} (e^{-t/2^j} - e^{-2t/2^j}) = \lim_{m\to\infty} -e^{-2t/2^1} + e^{-t/2^m} = 1 - e^{-t}$$

due to massive cancellation.

Let p_j form the geometric sequence with the common ratio $1/2$, that is $p_j = 1/2^j$.

Then the Poissonized variance can be calculated as follows:

$$V(t) = \lim_{m\to\infty} \sum_{j=1}^{m} (e^{-t/2^j} - e^{-2t/2^j}) = \lim_{m\to\infty} -e^{-2t/2^1} + e^{-t/2^m} = 1 - e^{-t}$$

due to massive cancellation. (The last term makes the main contribution!)

Let p_j form the geometric sequence with the common ratio $1/2$, that is $p_j = 1/2^j$.
Then the Poissonized variance can be calculated as follows:

$$V(t) = \lim_{m\to\infty} \sum_{j=1}^{m} (e^{-t/2^j} - e^{-2t/2^j}) = \lim_{m\to\infty} -e^{-2t/2^1} + e^{-t/2^m} = 1 - e^{-t}$$

due to massive cancellation. (The last term makes the main contribution!)
Hence $V(t) \to 1$ as $t \to \infty$ in accordance with Thm. 2 because the lagged ratio $p_{j+1}/p_j = 1/2$.

Let the common ratio be $1/4$, that is $p_j = 3/4^j$.
Then the Poissonized variance can be calculated as follows:

$$V(t) = \lim_{m\to\infty} v_m(t); \qquad v_m(t) = \sum_{j=1}^{m} (e^{-3t/4^j} - e^{-6t/4^j}).$$

Partial sums $v_m(t)$ satisfy the recursion

$$v_{m+1}(2t) = -e^{-3t} - v_m(t) + e^{-6t/4^{m+1}}.$$

Let the common ratio be $1/4$, that is $p_j = 3/4^j$.
Then the Poissonized variance can be calculated as follows:

$$V(t) = \lim_{m\to\infty} v_m(t); \qquad v_m(t) = \sum_{j=1}^{m} (e^{-3t/4^j} - e^{-6t/4^j}).$$

Partial sums $v_m(t)$ satisfy the recursion

$$v_{m+1}(2t) = -e^{-3t} - v_m(t) + e^{-6t/4^{m+1}}.$$

Take t and $m = m(t)$ sufficiently large, then $e^{-6t/4^{m+1}} \simeq 1$, e^{-3t} is small. If $t_j = \frac{4^j}{3}\ln 2$ then j-th summand $(e^{-3t/4^j} - e^{-6t/4^j}) = 1/4$ is maximal. Summation of 5 summands around it gives $v_m(t_j) > 0.501$ $(m \geq j+2)$. Hence $v_m(2t_j) \simeq v_{m+1}(2t_j) < 0.499$ and $V(t)$ oscillates. Actual amplitude of the oscillation is about 0.028 in this case.

Archibald, Knopfmacher, Prodinger (2006): If $p_j = cq^j$, then

$$V_n = \log_{1/q} 2 + \delta_V(\log_{1/q} n) + o(1), \qquad n \to \infty,$$

where

$$\delta_V(x) = \delta_E(x + \log_{1/q} 2) - \delta_E(x)$$

and δ_E is periodic with period 1 and has zero mean.

Archibald, Knopfmacher, Prodinger (2006): If $p_j = cq^j$, then

$$V_n = \log_{1/q} 2 + \delta_V(\log_{1/q} n) + o(1), \qquad n \to \infty,$$

where

$$\delta_V(x) = \delta_E(x + \log_{1/q} 2) - \delta_E(x)$$

and δ_E is periodic with period 1 and has zero mean.
So V_n converges iff $\log_{1/q} 2$ is integer, and it is the limit.
Thm. 2 extends this to "asymptotically geometric probabilities".

Archibald, Knopfmacher, Prodinger (2006): If $p_j = cq^j$, then

$$V_n = \log_{1/q} 2 + \delta_V(\log_{1/q} n) + o(1), \qquad n \to \infty,$$

where

$$\delta_V(x) = \delta_E(x + \log_{1/q} 2) - \delta_E(x)$$

and δ_E is periodic with period 1 and has zero mean.
So V_n converges iff $\log_{1/q} 2$ is integer, and it is the limit.
Thm. 2 extends this to "asymptotically geometric probabilities".
Karlin (1967) erroneously claimed that the variance converges for any geometric probabilities. Our motivation for study this question was, in particular, in the necessity to puzzle out this contradiction. It turns out that Karlin's sufficient condition for the convergence of V_n is in fact necessary and sufficient, and produces the correct criteria $\log_{1/q} 2 \in \mathbb{Z}$.

Suppose that the following regular variation assumption holds: for $y > 0$

$$\max\{j : p_j \geq 1/y\} \sim y^{\gamma}\ell(y), \qquad y \to \infty,$$

where $0 < \gamma \leq 1$ and ℓ is a slowly varying function.
(This case was considered by Karlin (1967)).
Then the inequality $p_{j+k(j)}/p_j \leq 2/3$ implies $k(j) \to \infty$ as $j \to \infty$.
So $\liminf \frac{p_{j+k}}{p_j} \geq \frac{1}{2}$ for any fixed k and Thm. 1 implies that $V_n \to \infty$ and K_n has asymptotically normal distribution.

We prove the statements for Poissonized version of the process, and then show how de-Poissonization can be done.
It is convenient to introduce the counting measure

$$\nu(dx) = \sum\nolimits_j \delta_{p_j}(dx)$$

and the function

$$\Delta\nu(x) = \nu((x/2, x]) = \#\{j : x/2 < p_j \leq x\}.$$

We prove the statements for Poissonized version of the process, and then show how de-Poissonization can be done.
It is convenient to introduce the counting measure

$$\nu(dx) = \sum\nolimits_j \delta_{p_j}(dx)$$

and the function

$$\Delta\nu(x) = \nu((x/2, x]) = \#\{j : x/2 < p_j \le x\}.$$

It turns out that bounds on "lagged ratio" p_{j+k}/p_j and variance of $K(t)$ can be expressed in terms of $\Delta\nu(x)$ for small x, providing an easy way to establish connections between them.

Lemma 1. For a fixed integer $k \geq 1$ the bound

$$\Delta\nu(x) \leq k$$

for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2}$$

holds for sufficiently large j.

Lemma 1. For a fixed integer $k \geq 1$ the bound

$$\Delta\nu(x) \leq k \qquad\qquad (\Delta\nu(x) \geq k)$$

for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \qquad\qquad \left(\frac{p_{j+k}}{p_j} \geq \frac{1}{2}\right)$$

holds for sufficiently large j.

Lemma 1. For a fixed integer $k \geq 1$ the bound

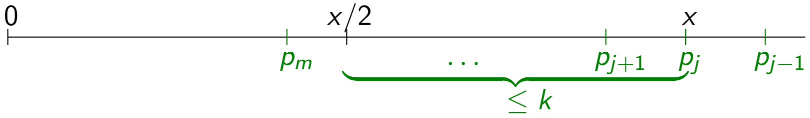
$$\Delta\nu(x) \leq k \qquad\qquad (\Delta\nu(x) \geq k)$$

for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \qquad\qquad \left(\frac{p_{j+k}}{p_j} \geq \frac{1}{2}\right)$$

holds for sufficiently large j.
Proof.

$\Delta\nu(x) \leq k \;\Rightarrow\; 2p_{j+k} \leq p_j$. Take $x = p_j$:



Lemma 1. For a fixed integer $k \geq 1$ the bound

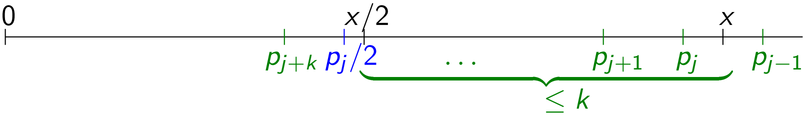
$$\Delta\nu(x) \leq k \qquad\qquad (\Delta\nu(x) \geq k)$$

for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \qquad\qquad \left(\frac{p_{j+k}}{p_j} \geq \frac{1}{2}\right)$$

holds for sufficiently large j.
Proof.

$2p_{j+k} \leq p_j \;\Rightarrow\; \Delta\nu(x) \leq k$. Let $p_j \leq x < p_{j-1}$



Lemma 1. For a fixed integer $k \geq 1$ the bound

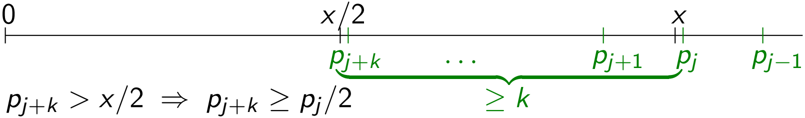
$$\Delta\nu(x) \leq k \qquad\qquad (\Delta\nu(x) \geq k)$$

for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \qquad\qquad \left(\frac{p_{j+k}}{p_j} \geq \frac{1}{2}\right)$$

holds for sufficiently large j.
Proof.

$\Delta\nu(x) \geq k \;\Rightarrow\; p_{j+k} \geq p_j/2$. Let $p_{j+1} \leq x < p_j$



$p_{j+k} > x/2 \;\Rightarrow\; p_{j+k} \geq p_j/2$

Lemma 1. For a fixed integer $k \geq 1$ the bound

$$\Delta\nu(x) \leq k \qquad\qquad (\Delta\nu(x) \geq k)$$

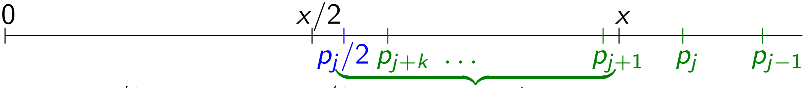
for sufficiently small $x > 0$ holds if and only if

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \qquad\qquad \left(\frac{p_{j+k}}{p_j} \geq \frac{1}{2}\right)$$

holds for sufficiently large j.

Proof.

$p_{j+k} \geq p_j/2 \;\Rightarrow\; \Delta\nu(x) \geq k$. Let $p_{j+1} \leq x < p_j$



$p_{j+k} > x/2 \;\Rightarrow\; p_{j+k} \geq p_j/2$

Lemma 2. The variance $V(t)$ can be represented as

$$V(t) = t \int_0^\infty e^{-tx} \Delta\nu(x)\, dx.$$

Lemma 2. The variance $V(t)$ can be represented as

$$V(t) = t \int_0^\infty e^{-tx} \Delta\nu(x)\, dx.$$

Proof. Recall the definition and rewrite it:

$$\Delta\nu(x) = \#\{j : x/2 < p_j \leq x\} = \sum\nolimits_j \mathbb{1}_{\{p_j \leq x < 2p_j\}}.$$

Lemma 2. The variance $V(t)$ can be represented as

$$V(t) = t \int_0^\infty e^{-tx} \Delta\nu(x)\, dx.$$

Proof. Recall the definition and rewrite it:

$$\Delta\nu(x) = \#\{j : x/2 < p_j \le x\} = \sum\nolimits_j \mathbb{1}_{\{p_j \le x < 2p_j\}}.$$

Hence

$$t \int_0^\infty e^{-tx} \Delta\nu(x)\, dx = t \int_0^\infty e^{-tx} \sum\nolimits_j \mathbb{1}_{\{p_j \le x < 2p_j\}}\, dx$$
$$= t \sum\nolimits_j \int_{p_j}^{2p_j} e^{-tx} dx = \sum\nolimits_j (e^{-tp_j} - e^{-2tp_j}) = V(t).$$

Proof of Thm. 1:

First part: $\limsup_j p_{j+k}/p_j \leq 1/2 \;\Rightarrow\; \overline{v} \leq k$.

- Suppose $k = 1$; otherwise divide the sequence (p_i) into k subsequences (p_{j+ki}), $j = 1, 2, \ldots, k$, and use additivity;

Proof of Thm. 1:

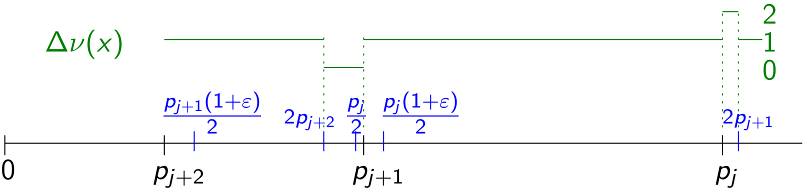
First part: $\limsup_j p_{j+k}/p_j \le 1/2 \ \Rightarrow\ \overline{v} \le k$.

- Suppose $k = 1$; otherwise divide the sequence (p_i) into k subsequences (p_{j+ki}), $j = 1, 2, \ldots, k$, and use additivity;
- Then $\Delta\nu(x) \le 2$ for x close to 0:
  for any $\varepsilon > 0$ $p_{j+1}/p_j \le (1+\varepsilon)/2$ for large j, and
  $p_{j+2}/p_j < 1/2$

Proof of Thm. 1:

First part: $\limsup_j p_{j+k}/p_j \le 1/2 \ \Rightarrow\ \overline{v} \le k$.

- Suppose $k = 1$; otherwise divide the sequence (p_i) into k subsequences (p_{j+ki}), $j = 1, 2, \ldots, k$, and use additivity;
- Then $\Delta\nu(x) \le 2$ for x close to 0:
  for any $\varepsilon > 0$ $p_{j+1}/p_j \le (1+\varepsilon)/2$ for large j, and
  $p_{j+2}/p_j < 1/2$



- Moreover, $\lambda(\{x \in [p_{j+1}, p_j] : \Delta\nu(x) = 2\}) \le \varepsilon p_j$.

- Let $D(x) = \int_0^x \Delta\nu(y)\,dy$. It is well defined, $D(0) = 0$ and $D(x) \le x + \varepsilon x$.

- Let $D(x) = \int_0^x \Delta\nu(y)\,dy$. It is well defined, $D(0) = 0$ and $D(x) \leq x + \varepsilon x$.
- Since $\varepsilon > 0$ is arbitrary, $\limsup_{x\downarrow 0} D(x)/x \leq 1$.

- Let $D(x) = \int_0^x \Delta\nu(y)\, dy$. It is well defined, $D(0) = 0$ and $D(x) \le x + \varepsilon x$.
- Since $\varepsilon > 0$ is arbitrary, $\limsup_{x \downarrow 0} D(x)/x \le 1$.
- Then integration by parts gives

$$\begin{aligned} V(t) &= t \int_0^\infty e^{-tx} dD(x) \\ &= t^2 \int_0^\infty e^{-tx} D(x)\, dx \\ &= \int_0^\infty y e^{-y} \frac{D(y/t)}{y/t} dy \end{aligned}$$

and by Fatou's lemma

$$\overline{v} = \limsup_{t \to \infty} V(t) \le \int_0^\infty y e^{-y} \limsup_{t \to \infty} \frac{D(y/t)}{y/t} dy \le 1.$$

Second part: $\overline{v} \le M \;\Rightarrow\; \exists k : \;\limsup \frac{p_{j+k}}{p_j} \le \frac{1}{2}$.
Due to the special structure:

$$V(t) = \sum_j (e^{-p_j t} - e^{-2p_j t}) \ge \sum_{p_j \in (x/2, x]} (e^{-p_j t} - e^{-2p_j t})$$
$$\ge \Delta\nu(x) \min_{p \in [x/2, x]} (e^{-pt} - e^{-2pt}).$$

Second part: $\overline{v} \le M \;\Rightarrow\; \exists k:\; \limsup \frac{p_{j+k}}{p_j} \le \frac{1}{2}$.
Due to the special structure:

$$V(t) = \sum_j (e^{-p_j t} - e^{-2p_j t}) \ge \sum_{p_j \in (x/2, x]} (e^{-p_j t} - e^{-2p_j t})$$
$$\ge \Delta\nu(x) \min_{p \in [x/2, x]} (e^{-pt} - e^{-2pt}).$$

Minimum is separated from zero: if $z = e^{-xt/2}$ then

$$\min_{p \in [x/2,x]} (e^{-pt} - e^{-2pt}) = \left\{ \begin{array}{ll} z^2 - z^4 & 0 \le z \le \frac{\sqrt{5}-1}{2} \\ z - z^2 & \frac{\sqrt{5}-1}{2} \le z \le 1 \end{array} \right\} \ge \sqrt{5} - 2 > 0.$$

Second part: $\overline{v} \le M \;\Rightarrow\; \exists k: \; \limsup \frac{p_{j+k}}{p_j} \le \frac{1}{2}$.
Due to the special structure:

$$V(t) = \sum_j (e^{-p_j t} - e^{-2p_j t}) \ge \sum_{p_j \in (x/2, x]} (e^{-p_j t} - e^{-2p_j t})$$
$$\ge \Delta\nu(x) \min_{p \in [x/2, x]} (e^{-pt} - e^{-2pt}).$$

Minimum is separated from zero: if $z = e^{-xt/2}$ then

$$\min_{p \in [x/2, x]} (e^{-pt} - e^{-2pt}) = \left\{ \begin{array}{ll} z^2 - z^4 & 0 \le z \le \frac{\sqrt{5}-1}{2} \\ z - z^2 & \frac{\sqrt{5}-1}{2} \le z \le 1 \end{array} \right\} \ge \sqrt{5} - 2 > 0.$$

Hence $\Delta\nu(x) \le \frac{2M}{\sqrt{5}-2}$ for small x, and the claim follows from Lemma 1.

Proof of Thm. 2:
Recall that $D(x) = \int_0^x \Delta\nu(x)dx$. Lemma 2 allows to write

$$V(t) = t\int_0^\infty e^{-tx}\Delta\nu(x)dx.$$

So $V(t) \to v \;\Leftrightarrow\; \int_0^\infty e^{-tx}dD(x) \sim v/t \;(t\to\infty)$.
By Karamata's Tauberian theorem this is equivalent to

$$\lim_{x\downarrow 0} D(x)/x = v. \qquad (*)$$

Proof of Thm. 2:
Recall that $D(x) = \int_0^x \Delta\nu(x)dx$. Lemma 2 allows to write

$$V(t) = t\int_0^\infty e^{-tx}\Delta\nu(x)dx.$$

So $V(t) \to v \;\Leftrightarrow\; \int_0^\infty e^{-tx}dD(x) \sim v/t \; (t \to \infty)$.
By Karamata's Tauberian theorem this is equivalent to

$$\lim_{x\downarrow 0} D(x)/x = v. \qquad (*)$$

Similarly as above $p_{j+k}/p_j \to 2$, $j \to \infty$ iff

$$\lim_{x\to 0}\frac{|u \in (0,x] : \Delta\nu(x) \neq k|}{x} = 0.$$

This is equivalent to (*).

Let $K_r(t) = \sum_j \mathbb{1}_{M_j(t)=r}$ be the number of values that occur exactly r times in the Poissonized model.
Its mean $\Phi_r(t) = \mathbb{E}[K_r(t)] = \frac{t^r}{r!} \int_0^\infty x^r e^{-tx} \nu(dx)$.

Let $K_r(t) = \sum_j \mathbb{1}_{M_j(t)=r}$ be the number of values that occur exactly r times in the Poissonized model.
Its mean $\Phi_r(t) = \mathbb{E}[K_r(t)] = \frac{t^r}{r!} \int_0^\infty x^r e^{-tx} \nu(dx)$.
The following estimates hold:

$$\Phi_n - \Phi(n) = O\left(\frac{\Phi_2(n)}{n}\right), \qquad V_n - V(n) \quad = O\left(\frac{\Phi_1(n)^2 + \Phi_2(n)}{n}\right).$$

Let $K_r(t) = \sum_j \mathbb{1}_{M_j(t)=r}$ be the number of values that occur exactly r times in the Poissonized model.
Its mean $\Phi_r(t) = \mathbb{E}[K_r(t)] = \frac{t^r}{r!} \int_0^\infty x^r e^{-tx} \nu(dx)$.
The following estimates hold:

$$\Phi_n - \Phi(n) = O\left(\frac{\Phi_2(n)}{n}\right), \qquad V_n - V(n) \; = O\left(\frac{\Phi_1(n)^2 + \Phi_2(n)}{n}\right).$$

The first follows from the inequality
$0 \le e^{-nx} - (1-x)^n \le nx^2 e^{-nx}$:

$$0 \le \Phi(n) - \Phi_n = \int_0^\infty (e^{-nx} - (1-x)^n)\nu(dx) \le \frac{2}{n}\Phi_2(n).$$

The second requires a bit more sophisticated but elementary analysis.
One can show that $\limsup \Phi_r(t) \le 2e\overline{v}$.

Thank you!