On the clustering of independent uniform random variables

Sándor Csörgő *
Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, Szeged, Hungary–6720 (csorgo@math.u-szeged.hu)

Wei Biao Wu
Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL 60637, U.S.A. (wbwu@galton.uchicago.edu)

ABSTRACT: We consider the number K_n of clusters at a distance level $d_n \in (0,1)$ of n independent random variables uniformly distributed in $[0,1]$, or the number K_n of connected components in the random interval graph generated by these variables and d_n, and, depending upon how fast $d_n \to 0$ as $n \to \infty$, determine the asymptotic distribution of K_n, with rates of convergence, and of related random variables that describe the cluster sizes.

KEYWORDS: Clusters of independent uniform random variables, number and size of clusters, asymptotic distributions, rates of convergence.

1. INTRODUCTION

Let $U_1, U_2, \ldots$ be independent random variables, each uniformly distributed in the unit interval $[0,1]$. For each $n \in \mathbb{N}$, let $U_{1,n} \le \cdots \le U_{n,n}$ be the order statistics pertaining to the sample $U_1, \ldots, U_n$. The elements of the sample are almost surely different, so that $U_{1,n} < \cdots < U_{n,n}$ almost surely. Given a deterministic threshold $d_n \in (0,1)$, the sequence $U_1, \ldots, U_n$ breaks up into nonempty disjoint clusters $C_{1,n}, ..., C_{K_n,n}$ at level d_n, where the random integer $K_n \in \{1, \ldots, n\}$ is the number of clusters, and we refer to the cardinality $N_{k,n} = |C_{k,n}|$, the number of elements in $C_{k,n}$, as the size or order of the cluster $C_{k,n}$, for which $\sum_{k=1}^{K_n} N_{k,n} = n$. Described in terms of spacings, this means that the set $\{U_1, \ldots, U_n\} = \{U_{1,n}, \ldots, U_{n,n}\} = \bigcup_{k=1}^{K_n} C_{k,n}$, where the distance between any two neighboring elements of $C_{k,n} = \{U_{N_{0,n}+\cdots+N_{k-1,n}+1,n}, \ldots, U_{N_{1,n}+\cdots+N_{k,n},n}\}$ is not greater than d_n, $k = 1, \ldots, K_n$, where $N_{0,n} = 0$, and, if $K_n > 1$ then $U_{N_{1,n}+\cdots+N_{k-1,n}+1,n} - U_{N_{1,n}+\cdots+N_{k-1,n},n} > d_n$, $k = 2, \ldots, K_n$, for the big spacings separating the clusters.

Now let $\mathcal{G}_n = \mathcal{G}(U_1, \ldots, U_n; d_n)$ be the random interval graph generated by the random variables $U_1, \ldots, U_n$ and the distance level d_n: the vertex set of $\mathcal{G}_n$ is the set $\{1, \ldots, n\}$, representing $U_1, \ldots, U_n$, such that there is an edge between the different vertices i and j, where $i, j \in \{1, \ldots, n\}$, if and only if $|U_i - U_j| \le d_n$, for which $\mathbb{P}\{|U_i - U_j| \le d_n\} = 2d_n - d_n^2$. In this language a cluster is a connected component $\mathcal{C}_{k,n}$ of $\mathcal{G}_n$ and the order $N_{k,n}$ of this cluster is the number of vertices in $\mathcal{C}_{k,n}$, so that $\mathcal{C}_{k,n}$ either consists of an isolated vertex or any two vertices of it are connected by a path of

[1] Research supported in part by the NSF Grant DMS–9625732 held at the University of Michigan and by the Hungarian National Foundation for Scientific Research, Grants T–032025 and T–034121.

edges, $k = 1, \dots, K_n$, and if the number of connected components $K_n > 1$ then there are no edges between any two clusters. (We use standard terminology as in [4].) Clearly, $\mathcal{G}_n = \mathcal{G}(U_1, \dots, U_n; d_n)$ is isomorphic to the random graph $\mathcal{G}(U_{1,n}, \dots, U_{n,n}; d_n)$.

More general random interval graphs, not necessarily based on the Uniform$[0,1]$ distribution, were considered to model clustering in [6] and [7], along with higher-dimensional analogues. Godehardt and Harris [7] obtained asymptotic Poisson distributions for the number of complete and maximal complete subgraphs of a fixed order and for the number of vertices of a fixed degree under specific conditions on the speed of $d_n \to 0$, assuming only the existence of an underlying density. However, refined results concerning asymptotic distributions for the number and size of the clusters are difficult to obtain without specifying the underlying distribution. Therefore, in the present paper we follow Godehardt and Jaworski [8], who continued the work in [7], in restricting ourselves to the uniform model above, which clearly is one of the most useful and natural one-dimensional models to understand some basic features. For further motivation and exposition of the area the we refer the reader to [3], [6]–[9], [11] and their references. Including extensions to higher dimensions, numerous related problems are investigated in the monographs by Hall [10], Aldous [1], Barbour, Holst and Janson [2], and Penrose [16], in the four-part survey by Hüsler [15] and in their vast number of references.

Godehardt and Jaworski [8] obtained numerous beautiful exact formulae in this Uniform(0,1) model, for example the one in their Theorem 1 stating that

$$\mathbb{P}\{K_n = k\} = \sum_{j=k-1}^{\min(n-1, \lfloor 1/d_n \rfloor)} (-1)^{k+j-1} \binom{n-1}{j} \binom{j}{k-1} (1 - jd_n)^n$$

for $k = 1, 2, \dots, \min(n-1, \lfloor 1/d_n \rfloor) + 1$, where $\lfloor x \rfloor = \max\{l \in \mathbb{Z} \colon l \le x\}$ is the integer part of $x \in \mathbb{R}$, and using these formulae and related other techniques they also derived several interesting asymptotic results. Aiming at all possible asymptotic distributions of K_n, described in the next section, this exact formula appears to be overly complicated: we use a technique based mainly on empirical distribution functions to obtain these results. Section 3 contains the results concerning the asymptotic behavior of cluster sizes.

All convergence relations are meant throughout as $n \to \infty$ unless otherwise specified. It is assumed that $d_n \to 0$. The results obtained in the paper may be transformed for the case when the underlying distribution is uniform on an arbitrary interval.

2. ASYMPTOTIC DISTRIBUTION OF THE NUMBER OF CLUSTERS

2.1. Results and discussion

Godehardt and Jaworski [8] show that when d_n is so small that $n^2 d_n \to 0$, then $n - K_n \to 0$ almost surely: there will only be clusters of size 1, or isolated vertices in $\mathcal{G}_n$. It also

follows from their results that if $n^2 d_n \to \lambda$ for some positive and finite constant λ, then $n - K_n = K_n(2) + o_{\mathbb{P}}(1)$, where $K_n(2)$ is the number of clusters of order 2, or the number of isolated edges in $\mathcal{G}_n$, and $K_n(2) \xrightarrow{\mathcal{D}} \mathcal{P}_\lambda$, where $\xrightarrow{\mathcal{D}}$ denotes convergence of distribution and $\mathcal{P}_\lambda$ stands for a Poisson random variable with mean λ.

The next meaningful case for K_n is when $nd_n \to 0$, but $n^2 d_n \to \infty$. In this case it is further assumed in [8] that $(nd_n)^{l-2} n^2 d_n = n^l d_n^{l-1} \to \lambda \in (0, \infty)$ for some $l \geq 3$, and shown for the number $K_n(l)$ of clusters of order l that $K_n(l) \xrightarrow{\mathcal{D}} \mathcal{P}_\lambda$ and that $K_n(m) \to 0$ almost surely for any $m > l$. Next, when $nd_n \to c \in (0, \infty)$ and J_n denotes the size of the cluster containing a given element of the sample $U_1, \ldots, U_n$, it is shown in [8] that $J_n + 1$ is asymptotically negative binomial of order 2 and parameter e^{-c}, and that clusters of the size greater than $\log n$ disappear. Third, when $nd_n \to \infty$ but $e^{nd_n}/n \to 0$, Godehardt and Jaworski [8] show that the limiting distribution of J_n/e^{nd_n} is Gamma with order 2 and parameter 1. Within this third case, they also prove that if $nd_n = \log \sqrt{nt_n}$ with $t_n \to t \in (0, \infty)$, then $K_n(m) \xrightarrow{\mathcal{D}} \mathcal{P}_{1/t}$ for each fixed $m \in \mathbb{N}$.

The overall number K_n of clusters is not treated in [8] in the range of d_n of the previous paragraph. Letting $\mathcal{N}(\mu, \sigma^2)$ denote a normal random variable with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, we prove that K_n is asymptotically normal in the whole range, but it turns out that this occurs in three different ways. Denoting by $\Phi(\cdot)$ the distribution function of $\mathcal{N}(0,1)$, we also derive rates of convergence in all three cases.

Theorem 2.1. *(i) If $nd_n \to 0$ and $n^2 d_n \to \infty$, then*

$$\begin{aligned}\Delta_n &:= \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}\,(1 - e^{-nd_n})}} \leq x \right\} - \Phi(x) \right| \\ &= O\left(\sqrt{\big[nd_n + \varepsilon_n\big] \log \frac{1}{nd_n}} + \frac{\log(n\sqrt{d_n}\,)}{n\sqrt{d_n}} \right),\end{aligned}$$

where $\varepsilon_n = \sqrt{(4 \log n)/n}$, and so $(K_n - ne^{-nd_n})/(n\sqrt{d_n}\,) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

(ii) If $0 < \liminf_{n\to\infty} nd_n \leq \limsup_{n\to\infty} nd_n < \infty$, then

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{K_n - ne^{-nd_n}}{\sqrt{n\, e^{-2nd_n}[e^{nd_n} - 1 - n^2 d_n^2]}} \leq x \right\} - \Phi(x) \right| = O\left(\frac{\log^{3/4} n}{n^{1/4}} \right),$$

and hence if $nd_n \to c \in (0, \infty)$, then $(K_n - ne^{-nd_n})\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, e^{-2c}\big[e^c - 1 - c^2\big]\big)$.

(iii) If $nd_n = \log(nr_n) \to \infty$, where $r_n = e^{nd_n}/n \to 0$, then

$$\Delta_n = O\left(\frac{\log^{3/2}(nr_n)}{\sqrt{nr_n}} + \sqrt{\varepsilon_n \log(nr_n) \log \frac{1}{r_n}} + \sqrt{r_n} \log \frac{1}{r_n} \right),$$

where Δ_n is as in case (i) and $\varepsilon_n = \sqrt{(4 \log n)/n}$ again, and so

$$\frac{K_n - ne^{-nd_n}}{\sqrt{n\, e^{-nd_n}}} = \frac{K_n - \frac{1}{r_n}}{\sqrt{1/r_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

It is interesting that the asymptotic variance is the same in cases (i) and (iii) while it assumes a different form in the middle case (ii). A referee noted that the mere asymptotic normality statements here could perhaps be obtained by the Poisson techniques for circular spacings in Section 7.2 of Barbour et al. [2], or directly derived from the central limit theorems there, which go back to Holst and Hüsler [14]. Even rates of convergence could be derived from their circular results, substantiating first Remark 7.2.1 in [2], at least for the extreme cases in (i) and (iii). Alternatively, our empirical-process method could be used to obtain convergence rates in the central limit results in Section 7.2 of [2].

A typical sequence $\{d_n\}$ for case (i) is $d_n = 1/n^\alpha$ for some $\alpha \in (1,2)$, in which case the resulting rate is $O\big(n^{-(\alpha-1)/2}\sqrt{\log n} + n^{-(2-\alpha)/2}\log n\big)$, which is fastest, namely $O\big(n^{-1/4}\log n\big)$, if $\alpha = 3/2$. Similarly, a typical sequence $\{r_n\}$ for case (iii) is $r_n = 1/n^\alpha$ for some $\alpha \in (0,1)$, when $d_n = (1-\alpha)(\log n)/n$, in which case the resulting rate in (iii) is $O\big(n^{-(1-\alpha)/2}\log^{3/2} n + n^{-1/4}\log^{5/4} n + n^{-\alpha/2}\log n\big)$, and this is fastest, $O\big(n^{-1/4}\log^{3/2} n\big)$, if $\alpha = 1/2$. With our method $O\big(n^{-1/4}\big)$, modulo logarithmic factors, is a natural limitation for the speed of convergence to normality; we believe it is in general.

The next order of magnitude for d_n is when r_n tends to a constant $r \in (0,\infty)$. In this case Theorem 12 of [8] states that $K_n - 1 \xrightarrow{\mathcal{D}} \mathcal{P}_{1/r}$, and this again could be obtained by the spacing techniques in [2]. Theorem 2.2 below strengthens this conclusion. We write $d_{\mathrm{TV}}(X,Y) = \sup\{|\mathbb{P}\{X \in B\} - \mathbb{P}\{Y \in B\}|\colon B \subset \{0,1,2\ldots\}\}$ for the total variation distance between the distributions of nonnegative integer-valued random variables X and Y ([2], pp. 1, 254), so that $d_{\mathrm{TV}}(X,Y) = \frac{1}{2}\sum_{k=0}^{\infty}\big|\mathbb{P}\{X=k\} - \mathbb{P}\{Y=k\}\big|$. Then we have

Theorem 2.2. *If $nd_n = \log(nr_n) \to \infty$, where $r_n = e^{nd_n}/n \to r \in (0,\infty)$, then*

$$\Delta_n^{(1)} := \max_{j\in\{0,1,2,\ldots\}}\big|\mathbb{P}\{K_n - 1 \le j\} - \mathbb{P}\{\mathcal{P}_{1/r_n} \le j\}\big| = O\bigg(\frac{\log^{3/2} n}{\sqrt{n}}\bigg)$$

and

$$\Delta_n^{(2)} := d_{\mathrm{TV}}\big(K_n - 1, \mathcal{P}_{1/r_n}\big) = O\bigg(\frac{\log^{5/2} n}{\sqrt{n}\,\log\log n}\bigg),$$

where the constants in the order bounds depend on r only, and $d_{\mathrm{TV}}(K_n - 1, \mathcal{P}_{1/r}) \to 0$.

Finally, when $r_n = e^{nd_n}/n \to \infty$, a result of Godehardt and Jaworski [8] rounds off the study, stating that $\mathbb{P}\{K_n = 1\} = \mathbb{P}\{\mathcal{G}_n \text{ is connected}\} \to 1$.

2.2. Proofs

Letting $Y_1, Y_2, \ldots$ denote a sequence of independent, identically exponentially distributed random variables with mean 1, so that $\mathbb{P}\{Y_1 > x\} = e^{-x}$ for all $x \ge 0$, with their partial sums $S_m = Y_1 + \cdots + Y_m$, $m \in \mathbb{N}$, the well-known distributional equality

$$\big(U_{1,n}, \ldots, U_{n,n}\big) \stackrel{\mathcal{D}}{=} \bigg(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\bigg), \quad n \in \mathbb{N},$$

then implies that $\mathcal{G}_n$ is isomorphic to the random graph $\mathcal{G}(S_1/S_{n+1},\dots,S_n/S_{n+1};d_n)$, or to $\mathcal{G}(S_1,\dots,S_n;d_nS_{n+1})$. Hence there is an edge between the vertices i and j of $\mathcal{G}_n$ if and only if $S_j-S_i=\sum_{l=i+1}^{j}Y_l\le d_nS_{n+1}$. But connectedness properties may be described by means of paths of edges of connecting vertices representing neighboring order statistics expressed by $S_1/S_{n+1},\dots,S_n/S_{n+1}$, and hence by the spacings $Y_2/S_{n+1},\dots,Y_n/S_{n+1}$. Indeed, for every $m=1,\dots,n$, $n\in\mathbb{N}$, it follows that

$$\begin{aligned}\mathbb{P}\{K_n=m\}&=\mathbb{P}\biggl\{\sum_{i=1}^{n-1}I\bigl\{S_{i+1}-S_i>d_nS_{n+1}\bigr\}=m-1\biggr\}\\&=\mathbb{P}\biggl\{\sum_{i=1}^{n-1}I\bigl\{Y_{i+1}>d_nS_{n+1}\bigr\}=m-1\biggr\},\end{aligned}$$

where $I\{A\}=I_A$ is the indicator of the event A, or, what is the same, $K_n\stackrel{\mathcal{D}}{=}1+\sum_{i=1}^{n-1}I\bigl\{Y_{i+1}>d_nS_{n+1}\bigr\}$, $n\in\mathbb{N}$. If we now introduce $F_n(x)=\frac{1}{n}\sum_{j=1}^{n}I\bigl\{Y_j\le x\bigr\}$, $x\in\mathbb{R}$, the empirical distribution function of $Y_1,\dots,Y_n$, then by the exchangeability of the sequence $Y_1,Y_2,\dots$ the last distributional equality implies

$$K_n\stackrel{\mathcal{D}}{=}n-(n-1)F_{n-1}\bigl(d_nS_{n+1}\bigr),\quad n=2,3,\dots,\tag{2.1}$$

and it also follows that

$$\mathbb{P}\{K_n\le k\}=\mathbb{P}\biggl\{\sum_{i=1}^{n-1}I\bigl\{Y_i>d_nS_{n+1}\bigr\}\le k-1\biggr\},\quad k=1,\dots,n\,.\tag{2.2}$$

Now let $G_n(t)=\frac{1}{n}\sum_{j=1}^{n}I\bigl\{U_j\le t\bigr\}$, $0\le t\le1$, be the uniform empirical distribution function. We state a special case of Lemma 2.3 of Stute [19] as

Lemma 2.1. *There exists a constant $x_*>0$ such that for all $0<\delta<1/8$ and $32\le s\le x_*\sqrt{\delta n}$ we have*

$$\mathbb{P}\biggl\{\sup_{0\le t\le\delta}\sqrt{n}\,\bigl|G_n(t)-t\bigr|>s\sqrt{\delta}\biggr\}\le4\,e^{-\frac{s^2}{16}}\,.$$

Proof of Theorem 2.1. Setting $F(x)=1-e^{-x}$ for $x\ge0$ and using (2.1), for every $n=2,3,\dots$ by elementary algebra we get

$$\begin{aligned}K_n-ne^{-nd_n}\stackrel{\mathcal{D}}{=}{}&n\bigl[e^{-d_nS_{n+1}}-e^{-nd_n}\bigr]-(n-1)\bigl[F_{n-1}\bigl(d_nS_{n+1}\bigr)-F\bigl(d_nS_{n+1}\bigr)\bigr]\\&+F\bigl(d_nS_{n+1}\bigr)\,.\end{aligned}\tag{2.3}$$

Introducing

$$\varepsilon_n=\frac{\sqrt{4\log n}}{\sqrt{n}}\,,\quad A_n=\biggl\{\biggl|\frac{S_{n+1}}{n+1}-1\biggr|\ge\varepsilon_n\biggr\}\quad\text{and}\quad q_n=\mathbb{P}\bigl\{A_n\bigr\},\tag{2.4}$$

Lemma 3.1 of Devroye [5] immediately implies that

$$q_n = \mathbb{P}\big\{A_n\big\} \le 2\, e^{-(n+1)\varepsilon_n^2/4} \le \frac{2}{n} \tag{2.5}$$

for all $n \ge 67$ for which $\varepsilon_n \le 1/2$. Also, with the complement A_n^c, for every $n = 2, 3, \ldots,$

$$\begin{aligned} I\{A_n^c\}\big|\big[F_{n-1}\big(d_n S_{n+1}\big) - F\big(d_n S_{n+1}\big)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big| \\ \le \sup_{(n+1)d_n(1-\varepsilon_n)\le t\le nd_n} \big|\big[F_{n-1}(t) - F(t)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big| \\ + \sup_{nd_n\le t\le (n+1)d_n(1+\varepsilon_n)} \big|\big[F_{n-1}(t) - F(t)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big|. \end{aligned} \tag{2.6}$$

Since the distributional equality $\{F_{n-1}(t) :\ t \in \mathbb{R}\} \stackrel{\mathcal{D}}{=} \{1 - G_{n-1}(e^{-t}) :\ t \in \mathbb{R}\}$ for all finite-dimensional distributions holds, we have

$$\begin{aligned} \sup_{nd_n(1-\varepsilon_n)\le t\le nd_n} \big|\big[F_{n-1}(t) - F(t)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big| \\ \stackrel{\mathcal{D}}{=} \sup_{(n+1)d_n(1-\varepsilon_n)\le t\le nd_n} \big|\big[G_{n-1}\big(e^{-t}\big) - e^{-t}\big] - \big[G_{n-1}\big(e^{-nd_n}\big) - e^{-nd_n}\big]\big|, \end{aligned}$$

and since $\{G_{n-1}(u) - G_{n-1}(v) : v \le u \le v + \delta\} \stackrel{\mathcal{D}}{=} \{G_{n-1}(u - v) : v \le u \le v + \delta\}$ for $0 \le v < v + \delta \le 1$, we obtain

$$\sup_{(n+1)d_n(1-\varepsilon_n)\le t\le nd_n} \big|\big[F_{n-1}(t) - F(t)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big| \stackrel{\mathcal{D}}{=} \Delta_n^-, \tag{2.7a}$$

where

$$\Delta_n^- = \sup_{0\le s\le \delta_n^-} |G_{n-1}(s) - s| \quad \text{with} \quad \delta_n^- = e^{-(n+1)d_n(1-\varepsilon_n)} - e^{-nd_n}.$$

Similarly,

$$\sup_{nd_n\le t\le (n+1)d_n(1+\varepsilon_n)} \big|\big[F_{n-1}(t) - F(t)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]\big| \stackrel{\mathcal{D}}{=} \Delta_n^+, \tag{2.7b}$$

where

$$\Delta_n^+ = \sup_{0\le s\le \delta_n^+} |G_{n-1}(s) - s| \quad \text{with} \quad \delta_n^+ = e^{-nd_n} - e^{-(n+1)d_n(1+\varepsilon_n)}.$$

Now the three cases (i), (iii) and (ii) are considered separately, in this order.

Case (i). We set $\sigma_n^2 = ne^{-nd_n}\big(1 - e^{-nd_n}\big)$, so that the asymptotic equality $\sigma_n^2 \sim n^2 d_n$ holds (meaning that the ratios of the two sides go to 1), and

$$u_n = 8\sqrt{nd_n \log \frac{1}{\min(1, nd_n)}}, \quad v_n = 64\sqrt{\varepsilon_n \log \frac{1}{\min(1, nd_n)}},$$

where ε_n is from the statement of the theorem and (2.4), and

$$w_n = \sqrt{n+1}\left[1 + \frac{1}{(n+1)d_n}\log\left(\frac{u_n\sigma_n}{n} + e^{-nd_n}\right)\right] \sim \frac{1}{\sqrt{n}\,d_n}\log\left(1 + e^{nd_n}\,\frac{u_n\sigma_n}{n}\right)$$
$$\sim 8\sqrt{\log\frac{1}{nd_n}}\,,$$

so that $w_n \to \infty$. Using (2.3), we decompose the random variable in question:

$$\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}\,(1-e^{-nd_n})}} = \frac{K_n - ne^{-nd_n}}{\sigma_n} = M_n^* + R_n^*\,, \tag{2.8}$$

where the main term is $M_n^* = -(n-1)[F_{n-1}(nd_n) - F(nd_n)]/\sigma_n$, while the remainder term is $R_n^* = R_n^{(1)} + R_n^{(2)} + R_n^{(3)}$ with

$$R_n^{(1)} = \frac{n\big[e^{-d_nS_{n+1}} - e^{-nd_n}\big]}{\sigma_n}\,, \quad 0 < R_n^{(3)} = \frac{F\big(d_nS_{n+1}\big)}{\sigma_n} < \frac{1}{\sigma_n} \quad \text{and}$$
$$R_n^{(2)} = -(n-1)\,\frac{\big[F_{n-1}\big(d_nS_{n+1}\big) - F\big(d_nS_{n+1}\big)\big] - \big[F_{n-1}(nd_n) - F(nd_n)\big]}{\sigma_n}\,.$$

Introducing $Z_{j,n} = \big[I\{Y_j \le nd_n\} - F(nd_n)\big]\big/\sqrt{e^{-nd_n}(1-e^{-nd_n})}$, we have

$$\begin{aligned}
\sup_{y\in\mathbb{R}}\big|\mathbb{P}\{M_n^* \le y\} - \Phi(y)\big| &\le \sup_{y\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\sum_{j=1}^{n-1} Z_{j,n}}{\sqrt{n-1}} \le y\sqrt{\frac{n}{n-1}}\right\} - \Phi\left(y\sqrt{\frac{n}{n-1}}\right)\right| \\
&\quad + \sup_{y\in\mathbb{R}}\left|\Phi\left(y\sqrt{\frac{n}{n-1}}\right) - \Phi(y)\right| \\
&\le \frac{D_1\sum_{j=1}^{n-1}\mathbb{E}(|Z_{j,n}|^3)}{\big[\sum_{j=1}^{n-1}\mathbb{E}\big(|Z_{j,n}|^2\big)\big]^{3/2}} + \frac{D_2}{n} \\
&= \frac{D_1}{\sqrt{n-1}}\,\frac{\mathbb{E}\big(|Z_{1,n}|^3\big)}{\big[\mathbb{E}\big(|Z_{1,n}|^2\big)\big]^{3/2}} + \frac{D_2}{n} \\
&\le \frac{D_1}{\sqrt{n-1}}\,\frac{1}{\sqrt{e^{-nd_n}(1-e^{-nd_n})}} + \frac{D_2}{n} \le \frac{D_3}{\sigma_n} \le \frac{D_4}{n\sqrt{d_n}}
\end{aligned} \tag{2.9}$$

by the Berry – Esseen theorem and elementary considerations, where $D_1, D_2, \ldots$ denote absolute constants. Also by the Berry – Esseen theorem, as applied to S_{n+1},

$$\mathbb{P}\big\{R_n^{(1)} > u_n\big\} = \mathbb{P}\left\{\sqrt{n+1}\left(1 - \frac{S_{n+1}}{n+1}\right) > w_n\right\} = \mathbb{P}\{\mathcal{N}(0,1) > w_n\} + O\left(\frac{1}{\sqrt{n}}\right)$$
$$= O\Big(e^{-w_n^2/3}\Big) + O\left(\frac{1}{\sqrt{n}}\right) = O\Big((nd_n)^{21}\Big) + O\left(\frac{1}{\sqrt{n}}\right).$$

Since $\mathbb{P}\big\{R_n^{(1)} < -u_n\big\}$ is of the same order, we have

$$\mathbb{P}\Big\{\Big|R_n^{(1)}\Big| > u_n\Big\} = O\Big((nd_n)^{21}\Big) + O\left(\frac{1}{\sqrt{n}}\right). \tag{2.10}$$

Next, for $n \ge 67$ we get by (2.4)–(2.7) that

$$\begin{aligned}\mathbb{P}\Big\{\Big|R_n^{(2)}\Big| > v_n\Big\} &\le \frac{2}{n} + \mathbb{P}\Big\{\Big|R_n^{(2)}\Big| > v_n\,,\, A_n^c\Big\} \\ &\le \frac{2}{n} + \mathbb{P}\bigg\{\sqrt{n-1}\,\Delta_n^- > \frac{v_n\sigma_n}{2\sqrt{n}}\bigg\} + \mathbb{P}\bigg\{\sqrt{n-1}\,\Delta_n^+ > \frac{v_n\sigma_n}{2\sqrt{n}}\bigg\}.\end{aligned}$$

Here, noticing that both $\delta_n^+ \sim nd_n\varepsilon_n$ and $\delta_n^- \sim nd_n\varepsilon_n$, we set

$$s_n^{\pm} = \frac{v_n\sigma_n}{2\sqrt{n\delta_n^{\pm}}} \sim \frac{v_n n\sqrt{d_n}}{2\sqrt{n}\sqrt{nd_n\varepsilon_n}} \sim 32\sqrt{\log\frac{1}{nd_n}}\,.$$

Fix any $x_* > 0$. For bounding each of the last two probabilities we separate the two possibilities $s_n^{\pm} \le x_*\sqrt{n\delta_n^{\pm}}$ or $s_n^{\pm} > x_*\sqrt{n\delta_n^{\pm}}$.

If $s_n^{\pm} \le x_*\sqrt{n\delta_n^{\pm}}$, then an application of Lemma 2.1 ensures that

$$\mathbb{P}\bigg\{\sqrt{n-1}\,\Delta_n^{\pm} > \frac{v_n\sigma_n}{2\sqrt{n}}\bigg\} \le 4\,e^{-(s_n^{\pm})^2/16} \le 4(nd_n)^{63}$$

for all n large enough within the first possibility.

If, on the other hand, $s_n^{\pm} > x_*\sqrt{n\delta_n^{\pm}}$, then we need to enlarge v_n a bit, putting

$$v_n^* := 64C\sqrt{\varepsilon_n\log\frac{1}{nd_n}} + 5\,\frac{\log(n\sqrt{d_n}\,)}{n\sqrt{d_n}} = C\,v_n + 5\,\frac{\log(n\sqrt{d_n}\,)}{n\sqrt{d_n}} \quad \text{and} \quad x_n = \frac{v_n^*\sigma_n}{2}\,\frac{\sqrt{n-1}}{\sqrt{n}}$$

with a constant $C = \max\{(3e/2x_*), 1\}$. Then we have

$$(n-1)\log\big(1+(e-1)\delta_n^{\pm}\big) < en\delta_n^{\pm} < \frac{ev_n\sigma_n}{2x_*} \le \frac{Cv_n\sigma_n}{3}\,,$$

and for all n large enough within the second possibility,

$$\begin{aligned}\mathbb{P}\bigg\{\sqrt{n-1}\,\Delta_n^{\pm} \ge \frac{v_n^*\sigma_n}{2\sqrt{n}}\bigg\} &\le \mathbb{P}\bigg\{\sqrt{n-1}\,\max\big\{\delta_n^{\pm}, G_{n-1}(\delta_n^{\pm})\big\} \ge \frac{v_n^*\sigma_n}{2\sqrt{n}}\bigg\} \\ &= \mathbb{P}\big\{(n-1)G_{n-1}(\delta_n^{\pm}) \ge x_n\big\} \\ &= \mathbb{P}\bigg\{\exp\bigg\{\sum_{j=1}^n I\big\{U_j \le \delta_n^{\pm}\big\}\bigg\} \ge e^{x_n}\bigg\} \\ &\le e^{-x_n}\mathbb{E}\bigg(\exp\bigg\{\sum_{j=1}^n I\big\{U_j \le \delta_n^{\pm}\big\}\bigg\}\bigg) \\ &= \exp\big\{-x_n + (n-1)\log\big(1+(e-1)\delta_n^{\pm}\big)\big\} \\ &< \exp\Big\{-2\log(n\sqrt{d_n}\,)\Big\} = \frac{1}{(n\sqrt{d_n}\,)^2}\,.\end{aligned} \tag{2.11}$$

Thus, combining the two possibilities,

$$\mathbb{P}\Big\{\Big|R_n^{(2)}\Big| > v_n^*\Big\} \le \frac{2}{n} + 8(nd_n)^{63} + \frac{2}{n^2d_n}\,, \tag{2.12}$$

and so, handling the trivial error term $R_n^{(3)}$ in an obvious fashion and collecting the bounds together from (2.10) and (2.12), for $t_n = u_n + v_n^* + \frac{1}{\sigma_n}$ we obtain

$$p_n^* = \mathbb{P}\big\{\big|R_n^*\big| > t_n\big\} = O\Big((nd_n)^{21}\Big) + O\bigg(\frac{1}{\sqrt{n}}\bigg) + O\bigg(\frac{1}{n^2 d_n}\bigg). \tag{2.13}$$

Using now the obvious inequality, resulting from (2.8),

$$\mathbb{P}\big\{M_n^* \le x - t_n\big\} - p_n^* \le \mathbb{P}\bigg\{\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-nd_n}\,(1-e^{-nd_n})}} \le x\bigg\} \le \mathbb{P}\big\{M_n^* \le x + t_n\big\} + p_n^*\,,$$

the inequality in (2.9) and the fact that

$$\sup_{x\in\mathbb{R}} \big|\Phi(x \pm t) - \Phi(x)\big| \le \frac{t}{\sqrt{2\pi}} \quad \text{for any } \ t \ge 0\,, \tag{2.14}$$

we obtain $\Delta_n = O(u_n + v_n^*) + O(1/\sigma_n) + p_n^*$, and the statement in (i) follows.

Case (iii). Using the decomposition in (2.8), the structure of the proof remains exactly the same as in case (i) if, keeping all other notation, we redefine

$$u_n = 8\,\frac{\log^{3/2}(nr_n)}{\sqrt{nr_n}}\,, \quad v_n = 64\sqrt{\varepsilon_n \log(nr_n) \log\frac{1}{r_n}} \quad \text{and} \quad v_n^* = C\sqrt{r_n}\,\log\frac{1}{r_n}\,,$$

where ε_n is as before and $C = 4{+}28(32)^2(e/x_*)$. Now, of course, $\sigma_n^2 \sim \sqrt{ne^{-nd_n}} = 1/r_n$. While formally the same with the new u_n, the asymptotic behavior of w_n now is

$$w_n = \sqrt{n+1}\,\bigg[1 + \frac{1}{(n+1)d_n}\log\bigg(\frac{u_n\sigma_n}{n} + e^{-nd_n}\bigg)\bigg] \sim \frac{1}{\sqrt{n}\,d_n}\log\bigg(1 + e^{nd_n}\,\frac{u_n\sigma_n}{n}\bigg)$$
$$\sim \frac{\log\big(1+u_n\sqrt{r_n}\big)}{\sqrt{n}\,d_n} = \frac{\log\Big(1+\frac{8\log^{3/2}(nr_n)}{\sqrt{n}}\Big)}{\sqrt{n}\,d_n} \sim \frac{8}{\sqrt{n}\,d_n}\,\frac{\log^{3/2}(nr_n)}{\sqrt{n}} \sim 8\sqrt{\log(nr_n)}\,,$$

so that $w_n \to \infty$ again.

Now, changing only the very last step, the argument in (2.9) yields

$$\sup_{y\in\mathbb{R}} \big|\mathbb{P}\{M_n^* \le y\} - \Phi(y)\big| \le \frac{D_3}{\sigma_n} \le D_4\sqrt{r_n}\,.$$

Also, with the modified u_n, the argument leading to (2.10) remains the same, now giving

$$\mathbb{P}\Big\{\Big|R_n^{(1)}\Big| > u_n\Big\} = O\bigg(\Big(\frac{1}{nr_n}\Big)^{21}\bigg) + O\bigg(\frac{1}{\sqrt{n}}\bigg).$$

Next, notice that $\delta_n^{\pm} \sim e^{-nd_n} nd_n\varepsilon_n = \varepsilon_n \log(nr_n)/(nr_n)$, and so

$$s_n^{\pm} = \frac{v_n\sigma_n}{2\sqrt{n\delta_n^{\pm}}} \sim \frac{32\sqrt{\varepsilon_n \log(nr_n)\log\frac{1}{r_n}}}{\sqrt{r_n}\sqrt{n}\sqrt{\frac{\varepsilon_n\log(nr_n)}{nr_n}}} = 32\sqrt{\log\frac{1}{r_n}}\,.$$

If $s_n \le x_* \sqrt{\delta_n^+ n}$, then by Lemma 2.1 again,

$$\mathbb{P}\Big\{ \sqrt{n-1}\, \Delta_n^{\pm} > \frac{v_n \sigma_n}{2\sqrt{n}} \Big\} \le 4\, e^{-(s_n^{\pm})^2/16} \le 4\, r_n^{63}$$

for all n large enough, while if $s_n > x_* \sqrt{\delta_n^+ n}$, then

$$(n-1) \log\big(1 + (e-1)\delta_n^{\pm}\big) < e n \delta_n^{\pm} < \frac{2e(32)^2}{x_*^2}\, \log\frac{1}{r_n} < \frac{28 e (32)^2}{C x_*^2}\, \frac{C\sigma_n \sqrt{r_n}}{7} \log\frac{1}{r_n} < \frac{\sigma_n v_n^*}{7}$$

and hence, with $x_n = v_n^* \sigma_n \sqrt{(n-1)/n}/2$ expressed in terms of the present v_n^*, by a simplified version of the argument in (2.11) for all n large enough we obtain

$$\begin{aligned} \mathbb{P}\Big\{ \sqrt{n-1}\, \Delta_n^{\pm} \ge \frac{v_n^* \sigma_n}{2\sqrt{n}} \Big\} &\le \exp\big\{ -x_n + (n-1)\log\big(1 + (e-1)\delta_n^{\pm}\big)\big\} \\ &\le \exp\Big\{ -\frac{\sigma_n v_n^*}{3} \Big\} \le \exp\Big\{ \frac{4\sigma_n \sqrt{r_n}\, \log r_n}{3} \Big\} \le r_n\,. \end{aligned}$$

Thus the argument leading to (2.12) this time produces $\mathbb{P}\big\{ \big| R_n^{(2)} \big| > v_n + v_n^* \big\} \le \frac{2}{n} + 8 r_n^{63} + 2 r_n$, and hence the analogue of (2.13) is

$$p_n^* = \mathbb{P}\big\{ \big| R_n^* \big| > t_n \big\} = O\bigg(\Big(\frac{1}{n r_n} \Big)^{21} + \frac{1}{\sqrt{n}} + r_n \bigg), \quad \text{where } t_n = u_n + v_n + v_n^* + \frac{1}{\sigma_n}\,.$$

So, substituting the present ingredients u_n, v_n, v_n^* and $1/\sigma_n \sim \sqrt{r_n}$ into the final equation $\Delta_n = O(u_n) + O(v_n) + O(v_n^*) + O(1/\sigma_n) + p_n^*$, case (iii) also follows.

Case (ii). The basic difference between the present "middle case" and the previous two "boundary cases" is that here $R_n^{(1)}$ is no longer a remainder term but, with a proper norming factor, it also contributes to the asymptotic distribution. This factor is presently redefined as the square root of $\sigma_n^2 = n\, e^{-2nd_n} \big[e^{nd_n} - 1 - n^2 d_n^2 \big] \sim e^{-2c} \big[e^c - 1 - c^2 \big] n$. Thus we need to modify the decomposition (2.8) for the present random variable of interest:

$$\frac{K_n - n e^{-nd_n}}{\sqrt{n\, e^{-2nd_n} [e^{nd_n} - 1 - n^2 d_n^2]}} = \frac{K_n - n e^{-nd_n}}{\sigma_n} = M_n^{\diamond} + R_n^{\diamond}\,, \tag{2.15}$$

where, introducing the independent and identically distributed random variables

$$V_{j,n} = \frac{n d_n e^{-nd_n}(1 - Y_j) - \big\{ I\{Y_j \le n d_n\} - F(n d_n) \big\}}{\sqrt{e^{-2nd_n} [e^{nd_n} - 1 - n^2 d_n^2]}}\,, \quad j = 1, \ldots, n+1\,,$$

where $\mathbb{E}(V_{j,n}) = 0$ and it can also be checked that $\mathbb{E}(V_{j,n}^2) = 1$, the main term now is

$$M_n^{\diamond} = \frac{n e^{-nd_n} \big[(n+1) d_n - d_n S_{n+1} \big] - (n+1) \big[F_{n+1}(n d_n) - F(n d_n) \big]}{\sigma_n} = \frac{\sum_{j=1}^{n+1} V_{j,n}}{\sqrt{n}}\,,$$

while the remainder term is $R_n^\diamond = \overline{R}_n^{(1)} + \overline{R}_n^{(2)} + \overline{R}_n^{(3)}$, where, from (2.3),

$$\overline{R}_n^{(1)} = \frac{ne^{-nd_n}\big[e^{nd_n - d_n S_{n+1}} - 1 - \{nd_n - d_n S_{n+1}\}\big]}{\sigma_n} = \frac{W_n^{(1)}}{\sigma_n}\,,$$

$$\overline{R}_n^{(2)} = (n-1)\,\frac{\big[F_{n-1}(nd_n) - F(nd_n)\big] - \big[F_{n-1}\big(d_n S_{n+1}\big) - F\big(d_n S_{n+1}\big)\big]}{\sigma_n} = (n-1)\,\frac{W_n^{(2)}}{\sigma_n}\,,$$

the latter formally agreeing with $R_n^{(2)}$ of cases (i) and (iii), but with a redefined σ_n, and

$$\overline{R}_n^{(3)} = \frac{F(d_n S_{n+1}) - nd_n e^{-nd_n} + \sum_{j=n}^{n+1}\big[I\{Y_j \le nd_n\} - F(nd_n)\big]}{\sigma_n}\,.$$

Going at it term by term, an obvious analogue of the argument in (2.9) now gives

$$\sup_{y\in\mathbb{R}}\big|\mathbb{P}\{M_n^\diamond \le y\} - \Phi(y)\big| = O\Big(\frac{1}{\sqrt{n}}\Big)\,. \tag{2.16}$$

Also, writing $c_n = nd_n$ and $\tau_n^2 = e^{-2c_n}\big[e^{c_n} - 1 - c_n^2\big]$, so that $\sigma_n = \tau_n\sqrt{n}$ and by assumption both sequences $\{c_n\}$ and $\{\tau_n\}$ are bounded away both from zero and infinity, setting the sequence u_n for the present case as

$$u_n = \frac{16c_n^2 e^{-c_n}}{\tau_n}\,\frac{\log n}{\sqrt{n}} = O\Big(\frac{\log n}{\sqrt{n}}\Big),$$

and using the notation in (2.4) and the inequality in (2.5), we obtain

$$\mathbb{P}\Big\{\Big|\overline{R}_n^{(1)}\Big| \ge u_n\Big\} \le \mathbb{P}\Big\{\Big|W_n^{(1)}\Big| \ge 16c_n^2 e^{-c_n}\,\log n\,,\,A_n^c\Big\} + \mathbb{P}\{A_n\} = \frac{2}{n}$$

for all n large enough, because if n is beyond some threshold, then on the event A_n^c we have $\big|W_n^{(1)}\big| < ne^{-c_n}(nd_n - d_n S_{n+1})^2 < ne^{-c_n}(2c_n\varepsilon_n)^2 < 16c_n^2 e^{-c_n}\log n$.

Next, keeping δ_n^- and δ_n^+ from (2.7) but redefining again s_n and v_n by setting $s_n = 32\sqrt{\log n}$ and $v_n = 2s_n\max\big(\sqrt{\delta_n^-}\,,\sqrt{\delta_n^+}\,\big)/\tau_n$, by (2.4)–(2.7) and Lemma 2.1,

$$\begin{aligned}
\mathbb{P}\Big\{\Big|\overline{R}_n^{(2)}\Big| \ge v_n\Big\} &= \mathbb{P}\Big\{\frac{(n-1)\big|W_n^{(2)}\big|}{\tau_n\sqrt{n}} \ge \frac{2s_n\max\big(\sqrt{\delta_n^-}\,,\sqrt{\delta_n^+}\,\big)}{\tau_n}\Big\} \\
&\le \mathbb{P}\Big\{\sqrt{n-1}\,\Big|W_n^{(2)}\Big| \ge 2s_n\max\Big(\sqrt{\delta_n^-}\,,\sqrt{\delta_n^+}\,\Big)\,,\,A_n^c\Big\} + \mathbb{P}\{A_n\} \\
&\le \mathbb{P}\Big\{\sqrt{n-1}\,\Delta_n^- \ge s_n\sqrt{\delta_n^-}\Big\} + \mathbb{P}\Big\{\sqrt{n-1}\,\Delta_n^+ \ge s_n\sqrt{\delta_n^+}\Big\} + \frac{2}{n} \\
&\le 8e^{-\frac{s_n^2}{16}} + \frac{2}{n} = \frac{8}{n^{64}} + \frac{2}{n} \le \frac{3}{n}
\end{aligned}$$

for all n large enough since in the present case $\delta_n^\pm \sim c_n e^{-c_n}\varepsilon_n$, and so the inequality $32 \le s_n \le x_*\sqrt{n\delta_n^\pm}$ is satisfied for all n large enough, regardless of the value of the constant x_* in Lemma 2.1. Note also that $v_n = O\big(\big[\log^{3/4} n\big]/n^{1/4}\big)$.

Since the error term $\overline{R}_n^{(3)}$ is again trivial, namely $\bigl|\overline{R}_n^{(3)}\bigr| \le D/\sqrt{n}$ for some constant $D > 0$, for the remainder term we altogether have

$$p_n^{\diamond} = \mathbb{P}\{|R_n^{\diamond}| > t_n\} = O\biggl(\frac{1}{n}\biggr), \quad \text{where } t_n = u_n + v_n + \frac{D}{\sqrt{n}} = O\biggl(\frac{\log^{3/4} n}{n^{1/4}}\biggr).$$

Putting this together with the inequality

$$\mathbb{P}\bigl\{M_n^{\diamond} \le x - t_n\bigr\} - p_n^{\diamond} \le \mathbb{P}\biggl\{\frac{K_n - ne^{-nd_n}}{\sqrt{ne^{-2nd_n}\,(e^{nd_n} - 1 - n^2 d_n^2)}} \le x\biggr\} \le \mathbb{P}\bigl\{M_n^{\diamond} \le x + t_n\bigr\} + p_n^{\diamond},$$

itself coming from (2.15), the bound in (2.16) for the main term and the inequality in (2.14), we see that the statement for the maximal deviation in case (ii) also follows. ∎

The proof of Theorem 2.2 requires the following

Lemma 2.2. *If $0 < \lambda < \mu$, then*

$$d_{\mathrm{TV}}\bigl(\mathcal{P}_\lambda, \mathcal{P}_\mu\bigr) \le \frac{\lfloor\lambda\rfloor^{\lfloor\lambda\rfloor}}{\lfloor\lambda\rfloor!}\, e^{-\lfloor\lambda\rfloor}\,(\mu - \lambda) \le \min\biggl(1, \frac{1}{\sqrt{\lambda}}\biggr)(\mu - \lambda).$$

Proof. Setting

$$\kappa = \min\biggl\{k \in \mathbb{N}\colon\ \frac{\mu^k}{k!}\, e^{-\mu} > \frac{\lambda^k}{k!}\, e^{-\lambda}\biggr\} - 1 = \biggl\lfloor \frac{\mu - \lambda}{\log\mu - \log\lambda}\biggr\rfloor \in \Bigl[\lfloor\lambda\rfloor, \lfloor\mu\rfloor\Bigr],$$

we obtain, with empty sums understood as zero, as before,

$$\begin{aligned}
\frac{1}{2}\sum_{k=0}^{\kappa}\biggl|\frac{\lambda^k}{k!}\, e^{-\lambda} - \frac{\mu^k}{k!}\, e^{-\mu}\biggr| &= \frac{1}{2}\sum_{k=0}^{\kappa}\biggl[\frac{\lambda^k}{k!}\, e^{-\lambda} - \frac{\mu^k}{k!}\, e^{-\mu}\biggr] \\
&= \frac{e^{-\lambda} - e^{-\mu}}{2} - \frac{1}{2}\sum_{k=1}^{\kappa}\int_\lambda^\mu \biggl[\frac{t^{k-1}}{(k-1)!}\, e^{-t} - \frac{t^k}{k!}\, e^{-t}\biggr]dt \\
&= \frac{e^{-\lambda} - e^{-\mu}}{2} - \frac{1}{2}\int_\lambda^\mu \biggl[e^{-t} - \frac{t^\kappa}{\kappa!}\, e^{-t}\biggr]dt = \frac{1}{2}\int_\lambda^\mu \frac{t^\kappa}{\kappa!}\, e^{-t}dt\,,
\end{aligned}$$

valid also in the case when $\kappa = 0$, and

$$\begin{aligned}
\frac{1}{2}\sum_{k=\kappa+1}^{\infty}\biggl|\frac{\mu^k}{k!}\, e^{-\mu} - \frac{\lambda^k}{k!}\, e^{-\lambda}\biggr| &= \frac{1}{2}\sum_{k=\kappa+1}^{\infty}\biggl[\frac{\mu^k}{k!}\, e^{-\mu} - \frac{\lambda^k}{k!}\, e^{-\lambda}\biggr] \\
&= \frac{1}{2}\sum_{k=\kappa+1}^{\infty}\int_\lambda^\mu \biggl[\frac{t^{k-1}}{(k-1)!}\, e^{-t} - \frac{t^k}{k!}\, e^{-t}\biggr]dt = \frac{1}{2}\int_\lambda^\mu \frac{t^\kappa}{\kappa!}\, e^{-t}dt\,,
\end{aligned}$$

whence

$$\begin{aligned}
d_{\mathrm{TV}}\bigl(\mathcal{P}_\lambda, \mathcal{P}_\mu\bigr) = \int_\lambda^\mu \frac{t^\kappa}{\kappa!}\, e^{-t}dt &\le (\mu - \lambda)\max_{\lambda \le t \le \mu}\frac{t^\kappa}{\kappa!}\, e^{-t} \le (\mu - \lambda)\,\frac{\kappa^\kappa}{\kappa!}\, e^{-\kappa} \\
&\le (\mu - \lambda)\,\frac{\lfloor\lambda\rfloor^{\lfloor\lambda\rfloor}}{\lfloor\lambda\rfloor!}\, e^{-\lfloor\lambda\rfloor} \le (\mu - \lambda)\min\biggl(1, \frac{1}{\sqrt{\lambda}}\biggr)
\end{aligned}$$

by elementary calculation. ∎

Proof of Theorem 2.2. Using the notation in (2.4), manipulation based on (2.2) gives

$$\mathbb{P}\biggl\{\sum_{i=1}^{n-1} I\bigl\{Y_i > d_n(n+1)(1+\varepsilon_n)\bigr\} \le k-1\biggr\} - q_n \le \mathbb{P}\{K_n \le k\} \le \mathbb{P}\biggl\{\sum_{i=1}^{n-1} I\bigl\{Y_i > d_n(n+1)(1-\varepsilon_n)\bigr\} \le k-1\biggr\} + q_n$$

for all $n \ge 2$ and $k \in \mathbb{N}$. Hence, with $\mathcal{B}_n^p$ denoting a Binomial(n,p) random variable,

$$\Delta_n^{(1)} \le 2\,\max\Bigl\{d_{\mathrm{TV}}\Bigl(\mathcal{B}_{n-1}^{p_n^-}, \mathcal{P}_{1/r_n}\Bigr), d_{\mathrm{TV}}\Bigl(\mathcal{B}_{n-1}^{p_n^+}, \mathcal{P}_{1/r_n}\Bigr)\Bigr\} + q_n\,,$$

where $p_n^{\pm} = \exp\{-d_n(n+1)(1\pm\varepsilon_n)\}$. Using the condition on $\{d_n\}$, it is easy to see that $p_n^{\pm} = O(n^{-1})$ and $\bigl|(n-1)p_n^{\pm} - r_n^{-1}\bigr| = O(\varepsilon_n \log n)$.

By a theorem of Prokhorov [18], as adjusted in [2], p. 2, there exists an absolute constant $C > 0$ such that $d_{\mathrm{TV}}(\mathcal{B}_n^p, \mathcal{P}_{np}) \le C\,p,\ 0 < p < 1$. Applying this with $p = p_n^{\pm}$, using Lemma 2.2, (2.5) and the bounds stated above, for all $n \ge 67$ we have

$$\begin{aligned}\Delta_n^{(1)} &\le 2\,d_{\mathrm{TV}}\Bigl(\mathcal{B}_{n-1}^{p_n^-}, \mathcal{P}_{(n-1)p_n^-}\Bigr) + 2\,d_{\mathrm{TV}}\bigl(\mathcal{P}_{(n-1)p_n^-}, \mathcal{P}_{1/r_n}\bigr)\\ &\quad + 2\,d_{\mathrm{TV}}\Bigl(\mathcal{B}_{n-1}^{p_n^+}, \mathcal{P}_{(n-1)p_n^+}\Bigr) + 2\,d_{\mathrm{TV}}\bigl(\mathcal{P}_{(n-1)p_n^+}, \mathcal{P}_{1/r_n}\bigr) + q_n\\ &\le 2\,C\,p_n^- + 2\,C\,p_n^+ + 2\biggl|(n-1)p_n^+ - \frac{1}{r_n}\biggr| + 2\biggl|(n-1)p_n^- - \frac{1}{r_n}\biggr| + \frac{2}{n} = O\Bigl(\frac{\log^{3/2} n}{\sqrt{n}}\Bigr),\end{aligned}$$

proving the first statement of the theorem.

For the proof of the second one, first note that by the de Moivre – Stirling formula

$$\mathbb{P}\bigl\{\mathcal{P}_{1/r_n} \ge \lceil \ell_n \rceil\bigr\} = e^{-1/r_n} \sum_{k=\lceil \ell_n \rceil}^{\infty} \frac{r_n^{-k}}{k!} \le C_r \sum_{k=\lceil \ell_n \rceil}^{\infty} e^{k - k\log(rk/2)} \le C_r \sum_{k=\lceil \ell_n \rceil}^{\infty} e^{-k\log((r\ell_n)/(2e))}$$

for some constant $C_r > 0$ and for all n large enough, where $\ell_n = (\log n)/(\log\log n)$ and $\lceil x \rceil = \min\{l \in \mathbb{Z}\colon\, l \ge x\}$ is the "upper integer part" of $x \in \mathbb{R}$. But the last sum is

$$\begin{aligned}\sum_{k=\lceil \ell_n \rceil}^{\infty} \Bigl(\frac{2e}{r\ell_n}\Bigr)^k &= \Bigl(\frac{2e}{r\ell_n}\Bigr)^{\lceil \ell_n \rceil} \frac{1}{1 - \frac{2e}{r\ell_n}} \le 2\Bigl(\frac{2e}{r\ell_n}\Bigr)^{\ell_n}\\ &= 2\exp\biggl\{-\biggl(1 - \frac{\log(2e/r)}{\log\log n} - \frac{\log\log\log n}{\log\log n}\biggr)\log n\biggr\} \le \frac{2}{n^{1-\varepsilon}}\end{aligned}$$

for any fixed $\varepsilon \in (0,1)$, for all n large enough, and hence by the first statement,

$$\mathbb{P}\bigl\{K_n - 1 \ge \lceil \ell_n \rceil\bigr\} = \mathbb{P}\bigl\{\mathcal{P}_{1/r_n} \ge \lceil \ell_n \rceil\bigr\} + O\Bigl(\frac{\log^{3/2} n}{\sqrt{n}}\Bigr) = O\Bigl(\frac{\log^{3/2} n}{\sqrt{n}}\Bigr).$$

Therefore,

$$\Delta_n^{(2)} \le \mathbb{P}\big\{K_n - 1 \ge \lceil \ell_n \rceil\big\} + \mathbb{P}\big\{\mathcal{P}_{1/r_n} \ge \lceil \ell_n \rceil\big\} + \sum_{k=0}^{\lceil \ell_n \rceil - 1} \big|\mathbb{P}\big\{K_n - 1 = k\big\} - \mathbb{P}\big\{\mathcal{P}_{1/r_n} = k\big\}\big|$$

$$\le \mathbb{P}\big\{K_n - 1 \ge \lceil \ell_n \rceil\big\} + \mathbb{P}\big\{\mathcal{P}_{1/r_n} \ge \lceil \ell_n \rceil\big\} + \sum_{k=0}^{\lceil \ell_n \rceil - 1} \big|\mathbb{P}\big\{K_n - 1 \le k\big\} - \mathbb{P}\big\{\mathcal{P}_{1/r_n} \le k\big\}\big|$$

$$+ \sum_{k=0}^{\lceil \ell_n \rceil - 1} \big|\mathbb{P}\big\{K_n - 1 \le k-1\big\} - \mathbb{P}\big\{\mathcal{P}_{1/r_n} \le k-1\big\}\big|$$

$$= O\Big(\frac{\log^{3/2} n}{\sqrt{n}}\Big) + \frac{\log n}{\log\log n}\, O\Big(\frac{\log^{3/2} n}{\sqrt{n}}\Big) = O\Big(\frac{\log^{5/2} n}{\sqrt{n}\,\log\log n}\Big),$$

proving the second statement. The third one follows from the second by Lemma 2.2. ■

3. ON THE ASYMPTOTIC DISTRIBUTION OF CLUSTER SIZES

3.1. Results and discussion

It is easy to see by the discussion leading to (2.1) that, given $K_n = k$, the vector of cluster sizes $(N_{1,n} \ldots, N_{k,n})$, satisfying $\sum_{i=1}^k N_i = n$, follows the Bose – Einstein distribution:

$$\mathbb{P}\big\{N_{1,n} = m_1, \ldots, N_{k,n} = m_k \,|\, K_n = k\big\} = \frac{I\big\{\sum_{i=1}^k m_i = n\big\}}{\binom{n-1}{k-1}} \tag{3.1}$$

for any sequence $m_1, \ldots, m_k, \ldots$ of positive integers and $k \in \{1, \ldots, n\}$, $n \in \mathbb{N}$. So, in comparison with J_n, mentioned at the beginning of Section 2, perhaps a more natural way to measure cluster size is to look at the number L_n of elements in a randomly chosen cluster: we choose at random one of the K_n clusters, each with the random probability $1/K_n$, and let $L_n = N_{R_n,n}$ be the number of its elements — or the number of vertices in the connected component of $\mathcal{G}_n$ chosen at random —, where $\mathbb{P}\{R_n = j \,|\, K_n = k\} = 1/k$, $j = 1, \ldots, k$ for every $k \in \{1, \ldots, n\}$. The first results for L_n are designed to be the companions of those for K_n in the three cases of Theorem 2.1.

Theorem 3.1. *(i) If $nd_n \to 0$ and $n^2 d_n \to \infty$, then $\mathbb{P}\{L_n = 1\} \to 1$.*
(ii) If $nd_n \to c \in (0, \infty)$, then $\mathbb{P}\{L_n = m\} \to e^{-c}(1-e^{-c})^{m-1}$ for each fixed $m \in \mathbb{N}$.
(iii) If $nd_n = \log(nr_n) \to \infty$, where $r_n = e^{nd_n}/n \to 0$, then

$$\mathbb{P}\bigg\{\frac{L_n}{e^{\,nd_n}} \le x\bigg\} = \mathbb{P}\bigg\{\frac{L_n}{nr_n} \le x\bigg\} \to 1 - e^{-x} \quad \text{for every} \quad x \ge 0\,.$$

The limiting geometric distribution with success probability e^{-c} in case (ii) will be obtained from the complete convergence in (3.3) below and the interesting equation

$$\mathbb{P}\{L_n = m\} = \mathbb{E}\Big(\frac{K_n(m)}{K_n}\Big), \quad m \in \{1, \ldots, n\}, \ n \in \mathbb{N}\,. \tag{3.2}$$

Consider also $\overline{M}_n = \max(N_{1,n}, \dots, N_{K_n,n})$ and $\underline{M}_n = \min(N_{1,n}, \dots, N_{K_n,n})$, the largest and the smallest cluster sizes. Since $\mathbb{P}\{\underline{M}_n = 1\} \geq \mathbb{P}\{L_n = 1\}$ for every $n \in \mathbb{N}$, under the conditions of Theorem 3.1(i) we of course have $\mathbb{P}\{\underline{M}_n = 1\} \to 1$, and some partial results for $\overline{M}_n$ may be derived from those in [8] reviewed at the beginning of Section 2 in subcases of case (i) in Theorems 2.1 and 3.1. The problem of the asymptotic behavior of both $\underline{M}_n$ and $\overline{M}_n$ is open in both cases (ii) and (iii) of Theorems 2.1 and 3.1. However, we can determine the asymptotic distribution of all three of L_n, $\overline{M}_n$ and $\underline{M}_n$ under the condition yielding the asymptotic Poisson behavior of K_n in Theorem 2.2.

Theorem 3.2. *If* $nd_n = \log(nr_n) \to \infty$, *where* $r_n = e^{nd_n}/n \to r \in (0, \infty)$, *then* $\mathbb{P}\{L_n \geq \lfloor nx \rfloor\} \to e^{-x/r}$, $\mathbb{P}\{\overline{M}_n \leq nx\} \to \overline{H}_r(x)$ *and* $\mathbb{P}\{\underline{M}_n \leq nx\} \to \underline{H}_r(x)$ *for every* $x \in (0,1)$, *where*

$$\overline{H}_r(x) = e^{-\frac{1}{r}} \sum_{k=\lfloor 1/x \rfloor + 1}^{\infty} \frac{\mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)}{r^{k-1}\sqrt{k}}, \quad \underline{H}_r(x) = 1 - e^{-\frac{1}{r}} \left[1 + \sum_{k=2}^{\lceil 1/x \rceil - 1} \frac{\mathrm{vol}_{k-1}\big(\underline{D}_k(x)\big)}{r^{k-1}\sqrt{k}} \right],$$

with an empty sum meant as zero, $\overline{D}_k(x) = \{(x_1, \dots, x_k) \in [0,x]^k \colon x_1 + \cdots + x_k = 1\}$, $\underline{D}_k(x) = \{(x_1, \dots, x_k) \in [x,1]^k \colon x_1 + \cdots + x_k = 1\}$, *and where* $\mathrm{vol}_{k-1}(\cdot)$ *stands for volume, the* $(k-1)$*-dimensional Lebesgue measure for every* $k \geq 2$.

The first statement implies that the limiting distribution function of L_n/n coincides with the exponential distribution function $1 - e^{-x/r}$ for $0 \leq x < 1$, but at $x = 1$ has a jump up to 1: this saltus of the size $e^{-1/r} = \lim_{n\to\infty} \mathbb{P}\{L_n = n\} = \lim_{n\to\infty} \mathbb{P}\{K_n = 1\}$ comes from Theorem 2.2. This implies $\mathbb{E}(L_n)/n \to e^{-1/r} + \frac{1}{r}\int_0^1 x e^{-x/r} dx = \int_0^1 e^{-x/r} dx = r(1 - e^{-1/r})$, and limiting formulae for higher-order moments can be obtained similarly. Noting that $\mathrm{vol}_{k-1}\big(\overline{D}_k(1)\big) = \sqrt{k}\,/(k-1)! = \mathrm{vol}_{k-1}\big(\underline{D}_k(0)\big)$ for every $k \geq 2$, the two jumps $\overline{H}_r(1) - \overline{H}_r(1-) = \underline{H}_r(1) - \underline{H}_r(1-) = e^{-1/r}$ are the same, from the same source as for L_n, and it is also interesting to notice that $\underline{H}_r(x) = 1 - e^{-1/r}$ for all $x \in [1/2, 1)$.

Returning to the middle case (ii) of Theorem 3.1, let $\overline{M}_n = N_n^{(1)} \geq \cdots \geq N_n^{(K_n)}$ be the decreasingly ordered cluster sizes $N_{1,n} \dots, N_{K_n,n}$. If $nd_n \to c \in (0, \infty)$, then Theorem 2 of Hill [13] directly implies that $N_n^{(k)}/\log n \xrightarrow{\mathbb{P}} 1/\log(1/[1 - e^{-c}])$ for each fixed $k \in \mathbb{N}$, where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability; note that the limit does not depend on k. On the other hand, it follows by Theorem 2.1(ii) that $K_n/n \xrightarrow{\mathbb{P}} e^{-c}$, and hence by Hill's [12] earlier theorem it follows for the proportion $K_n(m)/K_n$ of the number of clusters having any fixed size $m \in \mathbb{N}$ that $K_n(m)/K_n \xrightarrow{\mathbb{P}} e^{-c}(1 - e^{-c})^{m-1}$. Our last result strengthens both of these weak laws not only to almost sure convergence, but to certain exponential inequalities, which imply even complete convergence: $\sum_{n=1}^{\infty} \mathbb{P}\big\{\big|n^{-1}K_n - e^{-c}\big| \geq \varepsilon\big\} < \infty$ and, for each $m \in \mathbb{N}$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ \left| \frac{K_n(m)}{K_n} - e^{-c}(1 - e^{-c})^{m-1} \right| \geq \varepsilon \right\} < \infty \quad \text{for every } \varepsilon > 0. \tag{3.3}$$

Theorem 3.3. *If $nd_n \to c \in (0,\infty)$, then there exist functions $D_i\colon (0,\infty) \to (0,\infty)$, $i = 1,2,3$, such that*

$$\mathbb{P}\left\{\left|\frac{K_n}{n} - e^{-c}\right| \geq \varepsilon\right\} \leq 4\, e^{\frac{1}{1250}}\, e^{-D_1(c)\,\varepsilon^2\, n}, \quad 0 < \varepsilon \leq c_\diamond := \min\left(c, \frac{1}{10}\right),$$

and

$$\mathbb{P}\left\{\left|\sum_{m\in H}\left[\frac{K_n(m)}{K_n} - e^{-c}(1-e^{-c})^{m-1}\right]\right| \geq \varepsilon\right\} \leq D_2(c)\sqrt{n}\, e^{-D_3(c)\,\varepsilon^2\, n}, \quad 0 < \varepsilon \leq c_\diamond\,,$$

hold for all n large enough, where $H \subset \mathbb{N}$ is an arbitrary set.

3.2. Proofs

Proof of Theorem 3.1. Throughout we understand $\binom{m}{n} = 0$ if $m < n$. Concerning L_n, we clearly have $\mathbb{P}\{L_n = j \,|\, K_n = k\} = \mathbb{P}\{N_{R_n,n} = j \,|\, K_n = k\} = \mathbb{P}\{N_{1,n} = j \,|\, K_n = k\}$ for all $j,k \in \{1,\dots,n\}$. Since, given $K_n = k$, we have $N_{1,n} + \cdots + N_{k,n} = n$, and since among the $\binom{n-1}{k-1}$ vectors $(n_1,\dots,n_k)$ of positive integer solutions to the equation $n_1 + \cdots + n_k = n$ there are exactly $\binom{n-j-1}{k-2}$ vectors satisfying $n_1 = j$, we see that

$$\mathbb{P}\{L_n = j \,|\, K_n = k\} = \frac{\binom{n-j-1}{k-2}}{\binom{n-1}{k-1}}\,, \quad j,k \in \{1,\dots,n\}\,, \tag{3.4}$$

and

$$\mathbb{P}\{L_n \geq l \,|\, K_n = k\} = \sum_{j=l}^{n} \frac{\binom{n-j-1}{k-2}}{\binom{n-1}{k-1}} = \frac{\binom{n-l}{k-1}}{\binom{n-1}{k-1}}\,, \quad l,k \in \{1,\dots,n\}\,, \tag{3.5}$$

for every $n \in \mathbb{N}$. Now we turn to the separate cases.

Case (i). Theorem 2.1(i) implies that $K_n/n \xrightarrow{\mathbb{P}} 1$, and so, since $0 < K_n/n \leq 1$, by the moment convergence theorem also that $\mathbb{E}(K_n/n) \to 1$. Since

$$\mathbb{P}\{L_n = 1\} = \sum_{k=1}^{n} \frac{\binom{n-2}{k-2}}{\binom{n-1}{k-1}}\, \mathbb{P}\{K_n = k\} = \sum_{k=1}^{n} \frac{k-1}{n-1}\, \mathbb{P}\{K_n = k\} = \frac{n}{n-1}\left[\mathbb{E}\left(\frac{K_n}{n}\right) - \frac{1}{n}\right]$$

by (3.4), this establishes the first case.

Case (ii). Consider any $k,m \in \{1,\dots,n\}$, and let $k_n(1),\dots,k_n(n)$ be any sequence of nonnegative integers such that $k_n(1) + \cdots + k_n(n) = k$. Then it is easy to see that $\mathbb{P}\big\{L_n = m \,|\, K_n = k, K_n(1) = k_n(1),\dots,K_n(n) = k_n(n)\big\} = k_n(m)/k$ and hence also $\mathbb{P}\big\{L_n = m \,|\, K_n, K_n(1),\dots,K_n(n)\big\} = K_n(m)/K_n$. This implies $\mathbb{E}(K_n(m)/K_n) = \mathbb{E}(\mathbb{P}\{L_n = m \,|\, K_n, K_n(1),\dots,K_n(n)\}) = \mathbb{E}(\mathbb{E}(I\{L_n = m\} \,|\, K_n, K_n(1),\dots,K_n(n))) = \mathbb{E}(I\{L_n = m\}) = \mathbb{P}\{L_n = m\}$ for every $m \in \{1,\dots,n\}$, the equation highlighted in (3.2). Since for each fixed $m \in \mathbb{N}$ the bounded sequence $\{K_n(m)/K_n\}$ converges to $e^{-c}(1-e^{-c})^{m-1}$ almost surely by (3.3), the result in the second case also follows.

Case (iii). Using (3.4) and that $\binom{n-j-1}{k-2} = \binom{n-j}{k-1} - \binom{n-j-1}{k-1}$, for all $x > 0$ we obtain

$$\begin{aligned}
\mathbb{P}\{L_n \geq \lfloor nr_n x \rfloor\} &= \sum_{k=1}^{n} \sum_{j=\lfloor nr_n x \rfloor}^{n-(k-1)} \frac{\binom{n-j}{k-1} - \binom{n-j-1}{k-1}}{\binom{n-1}{k-1}} \mathbb{P}\{K_n = k\} \\
&= \sum_{k=1}^{n} \frac{\binom{n-\lfloor nr_n x \rfloor}{k-1}}{\binom{n-1}{k-1}} \mathbb{P}\{K_n = k\} = \sum_{k=1}^{n-\lfloor nr_n x \rfloor + 1} \frac{\binom{n-\lfloor nr_n x \rfloor}{k-1}}{\binom{n-1}{k-1}} \mathbb{P}\{K_n = k\} \\
&= \mathbb{E}\bigg(\frac{(n-K_n)(n-K_n-1)\cdots(n-K_n-\lfloor nr_n x \rfloor + 2)}{(n-1)(n-2)\cdots(n-\lfloor nr_n x \rfloor + 1)} \bigg) \\
&= \mathbb{E}\bigg(\bigg[1 - \frac{K_n - 1}{n-1}\bigg] \bigg[1 - \frac{K_n - 1}{n-2}\bigg] \cdots \bigg[1 - \frac{K_n - 1}{n - \lfloor nr_n x \rfloor + 1}\bigg] \bigg).
\end{aligned}$$

Introducing the event $B_n = \big\{ r_n^{-1} - r_n^{-2/3} \leq K_n \leq r_n^{-1} + r_n^{-2/3} \big\}$, Theorem 2.1(iii) implies that $\mathbb{P}\{B_n\} \to 1$. Hence by straightforward considerations,

$$\begin{aligned}
\mathbb{P}\{L_n \geq \lfloor nr_n x \rfloor\} &= \mathbb{E}\bigg(I_{B_n} \prod_{j=n-\lfloor nr_n x \rfloor + 1}^{n-1} \bigg[1 - \frac{K_n - 1}{j}\bigg] \bigg) + o(1) \\
&= \mathbb{E}\bigg(I_{B_n} \exp\bigg\{ \sum_{j=n-\lfloor nr_n x \rfloor + 1}^{n-1} \log\bigg(1 - \frac{K_n - 1}{j}\bigg) \bigg\} \bigg) + o(1) \\
&= \mathbb{E}\bigg(I_{B_n} \exp\bigg\{ - \sum_{j=n-\lfloor nr_n x \rfloor + 1}^{n-1} \bigg[\frac{K_n - 1}{j} + \theta_{j,n} \frac{(K_n-1)^2}{j^2} \bigg] \bigg\} \bigg) + o(1) \\
&= \mathbb{E}\bigg(I_{B_n} \exp\bigg\{ - (\lfloor nr_n x \rfloor - 1) \bigg[\frac{K_n - 1}{\xi_n} - \vartheta_n \frac{(K_n-1)^2}{\eta_n^2} \bigg] \bigg\} \bigg) + o(1)\,,
\end{aligned}$$

where $-2 \leq \theta_{j,n}, \vartheta_n \leq 2$ and $n - \lfloor nr_n x \rfloor + 1 \leq \xi_n, \eta_n \leq n-1$. The random variable within the last expectation is in $(0,1]$ and, since $r_n \to 0$, it is easy to see that it goes to e^{-x}. Using the bounded convergence theorem again, this proves the third case. ∎

Proof of Theorem 3.2. Consider first L_n. The case $x = 1$, as already pointed out, follows directly from Theorem 2.2, so we take $x \in (0,1)$. By (3.5) we have

$$\mathbb{P}\{L_n \geq \lfloor nx \rfloor\} = \sum_{k=1}^{n-\lfloor nx \rfloor + 1} f_{k,n}(x)\, \mathbb{P}\{K_n = k\} \quad \text{with} \quad f_{k,n}(x) = \frac{\binom{n-\lfloor nx \rfloor}{k-1}}{\binom{n-1}{k-1}}\,.$$

Clearly, $f_{k,n}(x) \to (1-x)^{k-1}$, and $\mathbb{P}\{K_n = k\} \to r^{-(k-1)} e^{-1/r}/(k-1)!$ by Theorem 2.2 for each fixed $k \in \mathbb{N}$. Also, elementary calculation shows that for every $\varepsilon > 0$ there exists a $k_* = k_*(\varepsilon, x) \in \mathbb{N}$ such that $\sup_{k > k_*} f_{k,n}(x) \leq \varepsilon$ for all n sufficiently large. Therefore,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}\{L_n \geq \lfloor nx \rfloor\} &\leq \sum_{k=1}^{k_*} \lim_{n\to\infty} f_{k,n}(x)\, \mathbb{P}\{K_n = k\} + \varepsilon \limsup_{n\to\infty} \sum_{k=k_*+1}^{\infty} \mathbb{P}\{K_n = k\} \\
&\leq \sum_{k=1}^{\infty} \frac{(1-x)^{k-1}}{r^{k-1}} \frac{e^{-1/r}}{(k-1)!} + \varepsilon = e^{-x/r} + \varepsilon.
\end{aligned}$$

Thus $e^{-x/r} = \sum_{k=1}^{\infty} \liminf_{n\to\infty} f_{k,n}(x)\,\mathbb{P}\{K_n = k\} \le \liminf_{n\to\infty} \mathbb{P}\{L_n \ge \lfloor nx\rfloor\} \le \limsup_{n\to\infty} \mathbb{P}\{L_n \ge \lfloor nx\rfloor\} \le e^{-x/r}$ by Fatou's lemma, completing the proof for L_n.

Consider again any $x \in (0,1)$. By (3.1), $\mathbb{P}\{\overline{M}_n \le nx\} = \sum_{k=2}^{\infty} \overline{g}_{k,n}(x)\,\mathbb{P}\{K_n = k\}$, where $\mathbb{P}\{K_n = k\} = 0$ for $k > n$ and

$$\begin{aligned}\overline{g}_{k,n}(x) &= \frac{\#\{(n_1,\dots,n_k) \in \mathbb{N}^k\colon\, n_1,\dots,n_k \le nx,\ \sum_{i=1}^k n_i = n\}}{\#\{(n_1,\dots,n_k) \in \mathbb{N}^k\colon\, \sum_{i=1}^k n_i = n\}} \\ &\to \frac{\mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)}{\mathrm{vol}_{k-1}\big(\overline{D}_k(1)\big)} = \frac{\mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)(k-1)!}{\sqrt{k}} =: \overline{g}_k(x)\end{aligned}$$

Using again Fatou's lemma and that $\mathbb{P}\{K_n = k\} \to r^{-(k-1)}e^{-1/r}/(k-1)!,\ k \in \mathbb{N}$,

$$\liminf_{n\to\infty} \mathbb{P}\{\overline{M}_n \le nx\} \ge \sum_{k=2}^{\infty} \liminf_{n\to\infty} \overline{g}_{k,n}(x)\,\mathbb{P}\{K_n = k\} = e^{-\frac{1}{r}} \sum_{k=2}^{\infty} \frac{\mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)}{r^{k-1}\sqrt{k}} = \overline{H}_r(x).$$

As to the upper bound, since $\overline{g}_{k,n}(x) \le 1$, for each fixed $l = 2, 3, \dots$ we see that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}\{\overline{M}_n \le nx\} &\le 1 + \sum_{k=2}^{l} \limsup_{n\to\infty} \big[\overline{g}_{k,n}(x) - 1\big]\,\mathbb{P}\{K_n = k\} \\ &= 1 + \sum_{k=2}^{l} \big[\overline{g}_k(x) - 1\big]\,\frac{e^{-1/r}}{r^{k-1}(k-1)!}\,,\end{aligned}$$

and, as $l \to \infty$, this converges to $\overline{H}_r(x) = e^{-\frac{1}{r}} \sum_{k=2}^{\infty} \mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)\big/\big(r^{k-1}\sqrt{k}\big) = e^{-\frac{1}{r}} \sum_{k=\lfloor 1/x\rfloor+1}^{\infty} \mathrm{vol}_{k-1}\big(\overline{D}_k(x)\big)\big/\big(r^{k-1}\sqrt{k}\big)$. The proof for $\mathbb{P}\{\underline{M}_n > x\}$ is analogous, the only difference is that $\underline{g}_{1,n}(x) \equiv 1 \equiv \underline{g}_1(x)$ for the corresponding functions. ∎

In preparation for the proof of the second statement of Theorem 3.3, consider for each $k \in \{1,\dots,n\}$ independent random variables $V_1^{k,n},\dots,V_k^{k,n}$ with a common geometric distribution with success probability k/n, so that $\mathbb{P}\{V_i^{k,n} = m\} = \frac{k}{n}\big(1-\frac{k}{n}\big)^{m-1}$, $m \in \mathbb{N}$, for all $i = 1,\dots,k$. Then, for any sequence $m_1,\dots,m_k$ of positive integers,

$$\begin{aligned}\mathbb{P}\left\{ \bigcap_{i=1}^k \{V_i^{k,n} = m_i\} \,\middle|\, \sum_{i=1}^k V_i^{k,n} = n \right\} &= \frac{\prod_{i=1}^k \mathbb{P}\{V_i^{k,n} = m_i\}}{\mathbb{P}\{\sum_{i=1}^k V_i^{k,n} = n\}}\, I\left\{ \sum_{i=1}^k m_i = n \right\} \\ &= \frac{\prod_{i=1}^k \frac{k}{n}\big(1-\frac{k}{n}\big)^{m_i-1}}{\binom{n-1}{k-1}\big(\frac{k}{n}\big)^k\big(1-\frac{k}{n}\big)^{n-k}}\, I\left\{ \sum_{i=1}^k m_i = n \right\} \\ &= \frac{I\{\sum_{i=1}^k m_i = n\}}{\binom{n-1}{k-1}}\,.\end{aligned}$$

Comparing with (3.1), the corresponding conditional distributions agree, in short:

$$\big(N_{1,n},\dots,N_{k,n} \,\big|\, K_n = k\big) \stackrel{\mathcal{D}}{=} \big(V_1^{k,n},\dots,V_k^{k,n} \,\big|\, V_1^{k,n} + \cdots + V_k^{k,n} = n\big) \qquad (3.6)$$

The following lemma is an analogue of Lemma 2.2 for geometric distributions.

Lemma 3.1. *If V_1 and V_2 are geometric random variables with respective success probabilities p_1 and p_2, where $0 < p_1 < p_2 < 1$, then*

$$d_{\mathrm{TV}}\big(V_1, V_2\big) \le \left(1 + \frac{1}{p_2}\right)\big(p_2 - p_1\big)\,.$$

Proof. Similarly as in the proof of Lemma 2.2, we set $\kappa = \min\big\{k \in \mathbb{N}\colon\ p_1 q_1^k > p_2 q_2^k\big\} - 1$, where $q_1 = 1 - p_1 > q_2 = 1 - p_2$, and hence $\big(\frac{q_1}{q_2}\big)^{\kappa+1} > \frac{p_2}{p_1}$ and $\big(\frac{q_1}{q_2}\big)^{\kappa} \le \frac{p_2}{p_1}$. Thus,

$$\begin{aligned}\frac{1}{2}\sum_{k=0}^{\kappa}\big|p_1 q_1^k - p_2 q_2^k\big| &= \frac{1}{2}\sum_{k=0}^{\kappa}\big[p_2 q_2^k - p_1 q_1^k\big] = \frac{p_2 - p_1}{2} - \frac{1}{2}\sum_{k=1}^{\kappa}\int_{q_2}^{q_1}\big[kt^{k-1} - (k+1)t^k\big]dt \\ &= \frac{p_2 - p_1}{2} - \frac{1}{2}\int_{q_2}^{q_1}\big[1 - (\kappa+1)t^{\kappa}\big]dt = \frac{q_1^{\kappa+1} - q_2^{\kappa+1}}{2}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{2}\sum_{k=\kappa+1}^{\infty}\big|p_1 q_1^k - p_2 q_2^k\big| &= \frac{1}{2}\sum_{k=\kappa+1}^{\infty}\big[p_1 q_1^k - p_2 q_2^k\big] \\ &= \frac{1}{2}\sum_{k=\kappa+1}^{\infty}\int_{q_2}^{q_1}\big[kt^{k-1} - (k+1)t^k\big]dt = \frac{q_1^{\kappa+1} - q_2^{\kappa+1}}{2}\,,\end{aligned}$$

whence, using the inequalities above,

$$\begin{aligned}d_{\mathrm{TV}}\big(V_1, V_2\big) &= q_1^{\kappa+1} - q_2^{\kappa+1} = \big[q_1^{\kappa+1} - q_1^{\kappa} q_2\big] + \big[q_1^{\kappa} q_2 - q_2^{\kappa+1}\big] \\ &\le \big[q_1 - q_2\big] + q_2^{\kappa+1}\left[\left(\frac{q_1}{q_2}\right)^{\kappa} - 1\right] \le \big[p_2 - p_1\big] + \frac{p_1 q_1^{\kappa+1}}{p_2}\left[\frac{p_2}{p_1} - 1\right],\end{aligned}$$

which implies the desired inequality. ∎

Proof of Theorem 3.3. First we consider the statement for $\{K_n\}$. Given $\varepsilon \in (0, c_\diamond]$, where $c_\diamond = \min(c, 1/10)$, put

$$\rho = \frac{\varepsilon}{2c} \quad \text{and} \quad D_{n,\rho} = \left\{\frac{S_{n+1}}{n+1} - 1 > -\rho\right\}.$$

Then by (2.1) and the half-sided version of (2.4),

$$\begin{aligned}\mathbb{P}\left\{\frac{K_n}{n} - e^{-c} \ge \varepsilon\right\} &\le \mathbb{P}\left\{\frac{n - (n-1)F_{n-1}(d_n S_{n+1})}{n} - e^{-c} \ge \varepsilon,\ \ D_{n,\rho}\right\} + \mathbb{P}\big\{D_{n,\rho}^c\big\} \\ &\le \mathbb{P}\Bigg\{\frac{n(1-\varepsilon) - ne^{-c}}{n-1} - F\big((1-\rho)(n+1)d_n\big) \\ &\qquad \ge F_{n-1}\big((1-\rho)(n+1)d_n\big) - F\big((1-\rho)(n+1)d_n\big)\Bigg\} + e^{-\rho^2 n/4} \\ &= \mathbb{P}\big\{F(u_n) - F_{n-1}(u_n) \ge v_n\big\} + e^{-\rho^2 n/4} \\ &= \mathbb{P}\left\{\sum_{j=1}^{n-1}\Big[\mathbb{E}\big(I\{Y_j \le u_n\}\big) - I\{Y_j \le u_n\}\Big] \ge (n-1)v_n\right\} + e^{-\rho^2 n/4}\,,\end{aligned}$$

where $u_n = (1-\rho)(n+1)d_n$ and

$$v_n = F\big((1-\rho)(n+1)d_n\big) - \frac{n(1-\varepsilon) - ne^{-c}}{n-1} = 1 - e^{-(1-\rho)(n+1)d_n} - \frac{n(1-\varepsilon) - ne^{-c}}{n-1}$$
$$\to \varepsilon - e^{-c}\big[e^{\varepsilon/2} - 1\big] > \varepsilon - \big[e^{\varepsilon/2} - 1\big] \geq \frac{2}{5}\,\varepsilon$$

since $c > 0$ and $0 < \varepsilon \leq c_\diamond \leq 1/10$. Thus by Hoeffding's inequality ([17], pp. 191–192),

$$\mathbb{P}\bigg\{\frac{K_n}{n} - e^{-c} \geq \varepsilon\bigg\} \leq e^{-2(2\varepsilon/5)^2(n-1)/4} + e^{-(\varepsilon^2 n)/(16c^2)} \leq 2\,e^{\frac{1}{1250}}\,e^{-D_1(c)\varepsilon^2 n}$$

for all n large enough, where $D_1(c) = \min\big(2/25, 1/(16c^2)\big)$. Since the deviation in the other direction only doubles the bound, the first statement of the theorem follows.

Turning to the second statement, let V be a geometric random variable with success probability e^{-c} and set $p_m = \mathbb{P}\{V = m\} = e^{-c}(1-e^{-c})^{m-1}$, $m \in \mathbb{N}$, and also $p_H = \mathbb{P}\{V \in H\} = \sum_{m\in H} p_m$ for the finite set $H \subset \mathbb{N}$ in the statement. Introduce also the set

$$H_n^\eta = \bigg\{k \in \{1,\dots,n\}\colon \bigg|\frac{k}{n} - e^{-c}\bigg| < \eta\bigg\}, \quad \text{where } \eta = \frac{e^{-c}}{4\big(1+e^{-c}+c\big)}\,\varepsilon\,.$$

Then, writing $p_n(k) = \mathbb{P}\{K_n = k\}$ for short, using (3.6) and then the first statement,

$$\begin{aligned}
p_n^H(\varepsilon) &:= \mathbb{P}\bigg\{\bigg|\sum_{m\in H}\bigg[\frac{K_n(m)}{K_n} - e^{-c}(1-e^{-c})^{m-1}\bigg]\bigg| \geq \varepsilon\bigg\} \\
&\leq \mathbb{P}\bigg\{\bigg|\sum_{m\in H}\frac{K_n(m)}{K_n} - p_H\bigg| \geq \varepsilon,\ \bigcup_{k\in H_n^\eta}\big\{K_n = k\big\}\bigg\} + \mathbb{P}\big\{K_n \notin H_n^\eta\big\} \\
&\leq \sum_{k\in H_n^\eta}\mathbb{P}\bigg\{\bigg|\frac{1}{k}\sum_{m\in H}K_n(m) - p_H\bigg| \geq \varepsilon\ \bigg|\ K_n = k\bigg\}\,p_n(k) + \mathbb{P}\big\{K_n \notin H_n^\eta\big\} \\
&= \sum_{k\in H_n^\eta}\mathbb{P}\bigg\{\bigg|\frac{1}{k}\sum_{i=1}^{k} I\big\{N_{i,n}\in H\big\} - p_H\bigg| \geq \varepsilon\ \bigg|\ K_n = k\bigg\}\,p_n(k) + \mathbb{P}\big\{K_n \notin H_n^\eta\big\} \\
&= \sum_{k\in H_n^\eta}\mathbb{P}\bigg\{\bigg|\frac{1}{k}\sum_{i=1}^{k} I\big\{V_i^{k,n}\in H\big\} - p_H\bigg| \geq \varepsilon\ \bigg|\ \sum_{i=1}^{k}V_i^{k,n} = n\bigg\}\,p_n(k) + \mathbb{P}\big\{K_n \notin H_n^\eta\big\} \\
&\leq \sum_{k\in H_n^\eta}\frac{\mathbb{P}\big\{\big|\frac{1}{k}\sum_{i=1}^{k} I\big\{V_i^{k,n}\in H\big\} - p_H\big| \geq \varepsilon\big\}}{\mathbb{P}\big\{\sum_{i=1}^{k}V_i^{k,n} = n\big\}}\,p_n(k) + 4\,e^{\frac{1}{1250}}\,e^{-D_1(c)\eta^2 n}
\end{aligned}$$

for all n large enough since $0 < \eta < \varepsilon \leq c_\diamond$.

The first term of this bound is not greater than

$$\bar{p}_n^H(\varepsilon) := \sum_{k\in H_n^\eta}\frac{\mathbb{P}\big\{\big|\frac{1}{k}\sum_{i=1}^{k} I\big\{V_i^{k,n}\in H\big\} - p_H^{k,n}\big| \geq \varepsilon - \big|p_H^{k,n} - p_H\big|\big\}}{\binom{n-1}{k-1}\big(\frac{k}{n}\big)^k\big(1-\frac{k}{n}\big)^{n-k}}\,p_n(k)\,,$$

where

$$p_H^{k,n} = \mathbb{E}\big(I\big\{V_1^{k,n} \in H\big\}\big) = \sum_{m\in H} \frac{k}{n}\bigg(1-\frac{k}{n}\bigg)^{m-1}, \quad k \in H_n^\eta\,.$$

Notice that for any $k \in H_n^\eta$,

$$\begin{aligned}\big|p_H^{k,n} - p_H\big| &\le \sum_{m\in H}\left|\frac{k}{n}\bigg(1-\frac{k}{n}\bigg)^{m-1} - e^{-c}(1-e^{-c})^{m-1}\right| \le 2\,d_{\mathrm{TV}}\big(V_1^{k,n}, V\big)\\ &\le 2\bigg(1+\frac{1}{\max\big(\frac{k}{n}, e^{-c}\big)}\bigg)\left|\frac{k}{n} - e^{-c}\right| < 2\bigg(1+\frac{1}{e^{-c}-\eta}\bigg)\eta < \frac{\varepsilon}{2}\end{aligned}$$

by Lemma 3.1, where the last inequality holds by the choice of η. Therefore, using this time the two-sided version of Hoeffding's inequality,

$$\bar p_n^H(\varepsilon) \le \sum_{k\in H_n^\eta} \frac{\mathbb{P}\big\{\big|\frac{1}{k}\sum_{i=1}^k I\big\{V_i^{k,n}\in H\big\} - p_H^{k,n}\big| \ge \frac{\varepsilon}{2}\big\} p_n(k)}{\binom{n-1}{k-1}\big(\frac{k}{n}\big)^k\big(1-\frac{k}{n}\big)^{n-k}} \le \sum_{k\in H_n^\eta} \frac{2e^{-\varepsilon^2 k/8}\,p_n(k)}{\binom{n-1}{k-1}\big(\frac{k}{n}\big)^k\big(1-\frac{k}{n}\big)^{n-k}}$$

where for all $k \in H_n^\eta$,

$$e^{-\varepsilon^2 k/8} < e^{-\varepsilon^2(e^{-c}-\eta)n/8} < e^{-D_4(c)\varepsilon^2 n} \quad \text{with} \quad D_4(c) = \frac{e^{-c}}{8}\bigg(1-\frac{1}{40(1+e^{-c}+c)}\bigg)$$

and, using the de Moivre – Stirling formula again,

$$\binom{n-1}{k-1}\bigg(\frac{k}{n}\bigg)^k\bigg(1-\frac{k}{n}\bigg)^{n-k} = \frac{k}{n}\,\frac{\sqrt{n}}{\sqrt{2\pi\frac{k}{n}\big(1-\frac{k}{n}\big)}}\,\big(1+o(1)\big)\,.$$

Thus, collecting the bounds, for all n large enough we have

$$p_n^H(\varepsilon) \le D_5(c)\sqrt{n}\,e^{-D_4(c)\varepsilon^2 n} + 4\,e^{\frac{1}{1250}}\,e^{-D_6(c)\varepsilon^2 n}$$

for some $D_5(c) > 0$ and $D_6(c) = e^{-2c}D_1(c)/[16(1+e^{-c}+c))^2]$, so the second statement follows with $D_2(c) = 2\max\big(D_5(c), 4\,e^{\frac{1}{1250}}\big)$ and $D_3(c) = \min\big(D_4(c), D_6(c)\big)$. ∎

ACKNOWLEDGEMENTS. We thank two referees for some useful remarks.

REFERENCES

[1] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer, New York, 1989.

[2] A.D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford University Press, Oxford, 1992.

[3] H.H. Bock, Probabilistic models in partitional cluster analysis, In: *Developments in Data Analysis: Metodološki Zvezki 12* (A. Ferligoj and A. Kramberger, eds.), FDV Ljubjana, Slovenia, 1996, pp. 3–25.

[4] B. Bollobás, *Graph Theory: An Introductory Course*, Springer, New York, 1979.

[5] L. Devroye, Laws of the iterated logarithm for order statistics of uniform spacings, *The Annals of Probability* **9** (1981), 860–867.

[6] E. Godehardt, *Graphs as Structural Models: The Application of Graphs and Multigraphs in Cluster Analysis*, Vieweg, Braunschweig, 1990.

[7] E. Godehardt and B. Harris, Asymptotic properties of random interval graphs and their use in cluster analysis, In: *Probabilistic Methods in Discrete Mathematics* (V.F. Kolchin, V.Ya. Kozlov, Yu.L. Pavlov and Yu.V. Prokhorov, eds.), VSP, Utrecht, 1997, pp. 19–30.

[8] E. Godehardt and J. Jaworski, On the connectivity of a random interval graph, *Random Structures and Algorithms* **9** (1996), 137–161.

[9] E. Godehardt, J. Jaworski and D. Godehardt, The application of random coincidence graphs for testing the homogeneity of data, In: *Classification, Data Analysis and Data Highways* (I. Balderjahn, R. Mathar and R. Sachader, eds.), Springer, Berlin, 1998, pp. 35–45.

[10] P. Hall, *Introduction to the Theory of Coverage Processes*, Wiley, New York, 1988.

[11] B. Harris and E. Godehardt, Probability models and limit theorems for random interval graphs with applications to cluster analysis, In: *Classification, Data Analysis and Data Highways* (I. Balderjahn, R. Mathar and R. Sachader, eds.), Springer, Berlin, 1998, pp. 54–61.

[12] B.M. Hill, Zipf's law and prior distributions for the composition of a population, *Journal of the American Statistical Association* **65** (1970), 1220–1232.

[13] B.M. Hill, The rank-frequency form of Zipf's law, *Journal of the American Statistical Association* **69** (1974), 1017–1026.

[14] L. Holst and J. Hüsler, On the random coverage of the circle, *Journal of Applied Probability* **21** (1984), 558–566.

[15] J. Hüsler, Coverage with uniform density in one dimension, on the circle; Coverage with uniform density in higher dimension; Coverage with nonuniform densities; Extension and related problems, *Rendiconti del Seminario Matematico di Messina. Serie II* **2**(**17**) (1993), *suppl.*, 25–40; 41–49; 51–63; 65–71.

[16] M. Penrose, *Random Geometric Graphs*, Oxford University Press, Oxford, 2003.

[17] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[18] Yu.V. Prokhorov, Asymptotic behavior of the binomial distribution [in Russian], *Uspekhi Matematičeskikh Nauk* **8** No.3 (55), 135–142.

[19] W. Stute, The oscillation behavior of empirical processes, *The Annals of Probability* **10** (1982), 86–107.