AIMS Mathematics, 11(11): 36134–36162.
DOI: 10.3934/math.20261400
Received: 13 September 2026
Revised: 20 October 2026
Accepted: 29 October 2026
Published: 03 November 2026

https://www.aimspress.com/journal/Math

Research article

A moment-matched negative binomial approximation for fiber intersection counts in two-dimensional random webs

Heuiju Chun[1] **and Jae-Hwan Jhong**[2,*]

[1] Department of Statistics and Information, Dongduk Women's University, 60 Hwarang-ro 13-gil, Seongbuk-gu, Seoul, 02748, Republic of Korea

[2] Department of Information Statistics, Chungbuk National University, 1 Chungdae-ro, Cheongju-si, Chungcheongbuk-do, 28644, Republic of Korea

* **Correspondence:** Email: jjh25@cbnu.ac.kr; Tel: +82432612256; Fax: +82432735928.

Abstract: The number of fiber intersections Y inside a bounded counting region of a two-dimensional random fiber web is a lattice-valued statistic whose finite-sample distribution can be asymmetric in a way that the recently established Gaussian limit cannot represent. That, together with the overdispersion of Y above a geometry-dependent fiber-count threshold and the mismatch between a continuous law and a lattice, motivates the negative binomial law matched to the exact finite-sample mean and variance of Y. We first showed that the conditional intersection probability depends on a partner fiber's seeding region only through its expected truncated length, so the moment formulas keep their form under type-specific seeding; this settles a question left open by Chun and Jhong (2026). We then provided the exact finite-sample threshold beyond which the moment-matched fit exists—weaker in our configuration than the leading-order condition by a factor of about 63—and settled which negative binomial parameter must diverge for asymptotic normality: Not r_n, for which we give a counterexample, but $r_n(1-\theta_n)$. A large-scale simulation then compared the two approximations on a common lattice, using closed-form truncated-length moments. Two corrections proved essential: The Kolmogorov–Smirnov statistic must be evaluated for a lattice reference law, and the normal law must carry a half-count continuity correction, which alone is worth some 60% of the apparent gain. The negative binomial's remaining advantage is real but not uniform: A factor of two in Kolmogorov distance at 150 equal-length fibers, the same at 300 on weaker evidence, not resolved from 500 onward, and reversed in one mixture configuration. On the Kolmogorov distance restricted to the outer 10% of the distribution, it was ahead in every configuration we ran, though not in every individual tail cell; at 150 fibers and three standard deviations, the corrected Gaussian understated the upper tail by a factor of 2.2 and the negative binomial by 15%.

Keywords: negative binomial approximation; overdispersion; random fiber web; geometric probability; count data; continuity correction
Mathematics Subject Classification: 60D05, 62E17, 62E20, 65C05

1. Introduction

Random fiber systems are canonical stochastic models for nonwoven materials, paper sheets, and fibrous composites in which fibers are idealized as line segments randomly placed in a planar region with random locations and orientations. The number of fiber intersections within a bounded observation window is a central descriptor of such systems, closely linked to strength, permeability, and acoustic behavior of the resulting materials [1–3].

Suh et al. [4] introduced a geometric probability model in which fibers are seeded with midpoints uniform over a seeding region S, and orientations are uniform on $[0, \pi)$, and intersections are counted within a bounded counting region $C \subset S$. Under an equal-length fiber model, they derived closed-form expressions for the expectation and variance of the intersection count Y. Chun [5] derived the exact distribution, and the first two moments, of the truncated fiber length L within C, providing the essential input for those moment formulas. Chun and Suh [6] extended the equal-length model to a two-component mixture of fiber lengths. More recently, Chun and Jhong [7] represented Y as a second-order U-statistic over unordered fiber pairs and applied Hoeffding's central limit theorem, proving that the appropriately normalized intersection count converges in distribution to a Gaussian limit in both models.

The Gaussian limit is a large-sample statement about a quantity that is, at every finite n, a nonnegative-integer-valued count. Two features of Y are invisible to it. Its support is a lattice, whereas the normal law is continuous and places mass on the negative half-line. In addition, the normal law is symmetric, whereas in the moderate-size configurations studied below, the simulated distribution of Y is positively skewed. (Whether Y is positively skewed at every finite n and for every admissible geometry is a question about its third cumulant, which we do not analyze; see Section 5.4.) Those are the two respects in which a finite-n description can improve on the limit, and they are what the negative binomial law offers: A lattice support and a positive skewness that its first two moments already determine.

Overdispersion enters differently, and it is worth separating the two points. Y is overdispersed—$\mathrm{Var}(Y) > \mathbb{E}(Y)$, because shared fibers induce positive local dependence among the pairwise intersection indicators Y_{ij}—but only above a threshold in n, which Proposition 1 identifies exactly. The situation is a familiar one in statistical modeling: An observed count or event whose law is shaped by a latent structure that is not itself recorded. Here, we have a sharing of fibers among pairs for which a tractable parametric family must be chosen to represent the law that structure induces—the mixture cure models of survival analysis, in which observed event times are governed by a latent cured-or-susceptible status, are a recent instance of the same template [8]. Overdispersion is the classical reason to prefer a negative binomial to a Poisson law; against a Gaussian, it is no argument at all, since a normal law can be given any variance one likes. What it does here is narrower and purely technical: It is the condition under which a negative binomial law with the moments of Y exists at all. Subject to it, approximating Y by $\mathrm{NB}(r, \theta)$ matched to its first two moments is the natural proposal, and one may expect an improvement at moderate n.

This paper develops that proposal and subjects it to a careful test. The moment matching itself (Lemma 3) is elementary algebra; the contributions lie in the seeding-region invariance, the exact threshold, the limit theorem for the negative binomial family under jointly varying parameters, and the design of the comparison. They are as follows.

1. **Seeding-region invariance.** (Section 2, Lemma 2) The conditional intersection probability given one fiber depends on the partner fiber's seeding region only through its expected truncated length. Consequently, the moment formulas of [4,6,7] are unchanged in *form* when the two fiber types are seeded over type-specific regions $S_1 \neq S_2$; only the numerical values of $\mathbb{E}(L_g)$, $\mathrm{Var}(L_g)$ change, and no additional geometric factor arises in the cross-type intersection probability. This resolves the question raised in [7, Remark 3.2] and removes a real ambiguity in how the mixture model is simulated.
2. **Moment-matched parameters and the exact overdispersion threshold.** (Section 3, Proposition 1) We derive the general principle $\theta = \mathbb{E}(Y)/\mathrm{Var}(Y)$, $r = \mathbb{E}(Y)^2/[\mathrm{Var}(Y) - \mathbb{E}(Y)]$, specialize it to both models, and record the exact finite-n condition $n > 2 + \mathbb{E}(L)^2/\{2\,\mathrm{Var}(L)\}$ under which the fit exists. For a square counting region of side 100 with $l = 15$, this reads $n > 4.08$, against $n \geq 256$ from the leading-order condition—a discrepancy that is entirely an artifact of discarding the Bernoulli diagonal term, which supplies 63% of $\mathrm{Var}(Y)$ at $n = 150$.
3. **A negative binomial central limit theorem under jointly varying parameters.** (Section 4, Proposition 3) Moment matching makes r_n and θ_n move together, so the relevant limit theorem is one in which both vary. It is not enough that $r_n \to \infty$, and we give a counterexample. The right condition is $r_n(1 - \theta_n) \to \infty$, which we show is both sufficient and necessary, together with the non-asymptotic bound $|\kappa_j(Z_n)| \leq (j-1)!\{r_n(1-\theta_n)\}^{-(j-2)/2}$ obtained from the Lévy measure of the negative binomial law. This makes precise the sense in which the negative binomial approximation nests, rather than contradicts, the Gaussian limit of [7].
4. **A fair simulation study.** (Section 5) We compare a continuity-corrected Gaussian and the moment-matched negative binomial on a common lattice, with Monte Carlo standard errors for the paired difference of the two fits, using replicate counts between 12000 and 40000 and closed-form truncated-length moments throughout. Two methodological points turn out to matter as much as the substantive one, and both act in the direction of exactly distorting this comparison; they are set out in Section 5.1. With both corrected, the negative binomial was the more accurate approximation at the smaller fiber counts of the equal-length model, and on the Kolmogorov distance restricted to the outer 10% of the distribution, it was ahead at every fiber count we ran; on the global criterion, the two were no longer distinguished from $n = 500$ onward, and in one mixture configuration, the corrected Gaussian was ahead.

The practical message is a qualified endorsement, and the qualification is as much a part of it as the endorsement. The numerical results support a regime-dependent use of the moment-matched negative binomial. Its improvements are clearest at the smaller fiber counts of the equal-length experiments and on the tail-based criteria—tail probabilities, extreme quantiles, and the confidence and prediction intervals built from them, which is where inference actually uses an approximation to a count. The advantage is not uniform across configurations or accuracy measures. Its global-distance advantage is not resolved at $n \geq 500$ in the equal-length configurations studied; the continuity-corrected Gaussian has the smaller global distance in the common-seeding mixture experiment, and at $n = 1200$, the fitted negative binomial has the larger error at the 99% quantile.

The two approximations are built from the same finite-n moments and share a Gaussian limit by Proposition 3; they differ in the third cumulant, which the normal law sets to zero and which the negative binomial, once its first two moments are fixed, has no freedom left to choose. Which of the two values is closer to the truth, and for how long, is a question about the third cumulant of Y itself. We

do not answer it here; what this paper supplies is the empirical answer for the configurations studied, obtained under conditions that make the comparison a fair one.

Section 2 fixes the model and proves the invariance lemma. Section 3 develops the moment-matched approximation and Section 4 its Gaussian limit. Section 5 reports the simulation study and Section 6 concludes the work.

2. Model, notation, and seeding-region invariance

2.1. Geometric setup

Let $C \subset \mathbb{R}^2$ be a bounded convex counting region of area $|C|$. A fiber of length l is generated by drawing its midpoint uniformly from a seeding region S and its orientation uniformly from $[0, \pi)$, independently. Throughout, we assume

$$S \;\supseteq\; C \oplus B(0, l/2), \tag{2.1}$$

where $\oplus$ denotes Minkowski addition, $A \oplus B = \{x + y : x \in A,\ y \in B\}$, and $B(0, \varrho) = \{x \in \mathbb{R}^2 : |x| \le \varrho\}$ is the closed disc of radius ϱ centered at the origin. Condition (2.1) guarantees that every fiber that can meet C has its midpoint in S. We write L for the length of the portion of a fiber lying inside C, the truncated length; because C is convex, this portion is a single segment, possibly empty. The exact distribution of L was obtained by Chun [5]. The following elementary identities are used repeatedly; part (ii) supplies, for the square regions used in Section 5, closed forms for exactly the truncated-length moments that the approximations of Section 3 require, so that no Monte Carlo error enters them.

Lemma 1. *(Moments of the truncated length) Assume that* (2.1) *holds. Then*

(i) $\mathbb{E}(L) = l\,|C|/|S|$, *whatever the shape of* C *and* S*;*
(ii) *if in addition* C *is a square of side* a *and* $l \le a$, *then for every integer* $k \ge 2$,

$$\mathbb{E}(L^k) = \frac{k(k-1)}{|S|} \int_0^l (l - t)\, t^{k-2} \Big(a^2 - \frac{4at}{\pi} + \frac{t^2}{\pi}\Big) dt, \tag{2.2}$$

and in particular

$$\mathbb{E}(L^2) = \frac{a^2 l^2 - \dfrac{4al^3}{3\pi} + \dfrac{l^4}{6\pi}}{|S|}, \qquad \mathbb{E}(L^3) = \frac{a^2 l^3 - \dfrac{2al^4}{\pi} + \dfrac{3l^5}{10\pi}}{|S|}. \tag{2.3}$$

Proof. Write the fiber as $\{m + t\,u(\vartheta) : |t| \le l/2\}$ with $u(\vartheta) = (\cos\vartheta, \sin\vartheta)$, $m \sim \mathrm{Unif}(S)$, and $\vartheta \sim \mathrm{Unif}[0, \pi)$, and represent L^k as a k-fold integral of indicators along the fiber:

$$L^k = \int_{[-l/2, l/2]^k} \prod_{i=1}^{k} \mathbf{1}\{m + t_i\,u(\vartheta) \in C\}\, dt_1 \cdots dt_k.$$

All k points lie in C exactly when m lies in $\bigcap_i \{C - t_i\,u(\vartheta)\}$, and since the shifts are collinear, this intersection is $\{C - t_{\min} u(\vartheta)\} \cap \{C - t_{\max} u(\vartheta)\}$, a set of area $A(t, \vartheta) := |C \cap \{C - t\,u(\vartheta)\}|$ with $t = t_{\max} - t_{\min}$. That set is contained in $C \oplus B(0, l/2) \subseteq S$, so its whole area is available and the seeding region enters

only through the density $1/|S|$. Taking $k = 1$ gives $A \equiv |C|$ and hence (i). For $k \geq 2$, the range $t_{\max} - t_{\min}$ of k points spread over an interval of length l contributes the factor $k(k-1)(l-t)t^{k-2}$, so

$$\mathbb{E}(L^k) = \frac{k(k-1)}{|S|} \int_0^l (l-t)\, t^{k-2}\, \bar{A}(t)\, dt, \qquad \bar{A}(t) = \frac{1}{\pi} \int_0^\pi A(t,\vartheta)\, d\vartheta.$$

For a square of side a and $t \leq a$, one has $A(t,\vartheta) = (a - t|\cos\vartheta|)(a - t|\sin\vartheta|)$, $\int_0^\pi |\cos\vartheta|\, d\vartheta = \int_0^\pi |\sin\vartheta|\, d\vartheta = 2$, and $\int_0^\pi |\cos\vartheta \sin\vartheta|\, d\vartheta = 1$, so that $\bar{A}(t) = a^2 - 4at/\pi + t^2/\pi$. This is (2.2), and evaluating the integrals at $k = 2, 3$ gives (2.3). □

The argument uses only the convexity of C: For any bounded convex C, the displayed formula for $\mathbb{E}(L^k)$ holds with $\bar{A}$ the isotropic covariogram of C, so the truncated-length moments reduce to a one-dimensional integral whenever $\bar{A}$ is available. For a rectangle of sides $a \times b$ and $t \leq \min(a, b)$, one has $\bar{A}(t) = ab - 2(a+b)t/\pi + t^2/\pi$, and for a disc of radius ϱ, $\bar{A}(t) = 2\varrho^2 \arccos\{t/(2\varrho)\} - \frac{t}{2}\sqrt{4\varrho^2 - t^2}$; both are used in Section 5.6.

2.2. Intersection count and seeding-region invariance

For n fibers, we write ξ_i for the marker (midpoint, orientation, length) of fiber i, let

$$Y_{ij} = \mathbf{1}\{\text{fibers } i \text{ and } j \text{ intersect within } C\}, \qquad Y = \sum_{1 \leq i < j \leq n} Y_{ij},$$

and let L_i be the truncated length of fiber i. The next lemma is the analytical engine of the whole theory. It is stated so as to make explicit that the two fibers may be seeded over *different* regions.

Lemma 2. *(Conditional intersection probability; seeding invariance) Fix fiber i with truncated length $L_i = t$, and let fiber j be generated independently with length l_j, with a midpoint uniform on a region S_j satisfying* (2.1) *for l_j, and an orientation uniform on* $[0, \pi)$*. Then*

$$\mathbb{P}(Y_{ij} = 1 \mid \xi_i) = \frac{2\, t\, l_j}{\pi |S_j|} = \frac{2\, t\, \mathbb{E}(L_j)}{\pi |C|}, \tag{2.4}$$

where $\mathbb{E}(L_j)$ is evaluated under S_j. In particular, $\mathbb{P}(Y_{ij} = 1) = 2\mathbb{E}(L_i)\mathbb{E}(L_j)/(\pi|C|)$, with each expectation being taken under its own seeding region.

Proof. The event $\{Y_{ij} = 1\}$ is exactly the event that the segment of fiber j crosses $I := (\text{fiber } i) \cap C$, a segment of length t and some direction α; any such crossing point lies in $I \subseteq C$. Fix the orientation φ of fiber j and write $J_\varphi = \{s\, u(\varphi) : |s| \leq l_j/2\}$. The set of midpoints m_j for which $m_j + J_\varphi$ meets I is $I \oplus J_\varphi$, a parallelogram of area $t\, l_j |\sin(\varphi - \alpha)|$; since $I \subseteq C$ and $J_\varphi \subseteq B(0, l_j/2)$, it is contained in $C \oplus B(0, l_j/2) \subseteq S_j$, so its entire area is available. Averaging over $\varphi \sim \text{Unif}[0, \pi)$,

$$\mathbb{P}(Y_{ij} = 1 \mid \xi_i) = \frac{1}{\pi |S_j|} \int_0^\pi t\, l_j |\sin(\varphi - \alpha)|\, d\varphi = \frac{2\, t\, l_j}{\pi |S_j|}.$$

Lemma 1 gives $l_j/|S_j| = \mathbb{E}(L_j)/|C|$, which is the second equality; the unconditional statement follows on averaging over L_i. □

Remark 1. *(Resolution of an open point) Remark 3.2 of Chun and Jhong [7] conjectured that allowing type-dependent seeding regions $S_1 \neq S_2$ in the two-length mixture model "introduces an extra geometric factor in the cross-type intersection probability", and left the matter as a topic for future work. Lemma 2 shows that this is not so: The factor $|S_j|$ cancels exactly against the definition of $\mathbb{E}(L_j)$, so the cross-type probability retains the form $2\mathbb{E}(L_1)\mathbb{E}(L_2)/(\pi|C|)$ whatever the seeding regions, provided only that each satisfies* (2.1)*. Since* (2.4) *is linear in t, every covariance computation in [4, 6, 7]—each of which proceeds by conditioning on a shared fiber—goes through verbatim. The exact mean and variance formulas, and the central limit theorems, therefore hold under either convention, with $\mathbb{E}(L_g), \mathrm{Var}(L_g)$ evaluated under S_g. This is verified numerically in Section 5.5. We stress that the two conventions are nevertheless* different probability models*, producing mean intersection counts that differ by roughly a quarter in the baseline configuration below, and we state explicitly which convention each experiment uses.*

Two remarks on the scope of Lemma 2 are in order. First, it is a statement about one ordered pair of fibers, so it extends without change to any number of fiber types and to fibers whose lengths are drawn from a general distribution: The conditional probability is $2t\,\mathbb{E}(L_j)/(\pi|C|)$ whatever the law of fiber j, provided its seeding region satisfies (2.1) for every length it can take. Second, the proof uses two things, uniform seeding together with (2.1), which make the whole parallelogram $I \oplus J_\varphi$ available with density $1/|S_j|$, and uniform orientation, which averages $|\sin(\varphi - \alpha)|$ to $2/\pi$ independently of the direction α of fiber i. If the seeding is non-uniform, or (2.1) fails so that part of $I \oplus J_\varphi$ lies outside S_j, the shape of the seeding region re-enters the probability; if the orientation law is not uniform, the average of $|\sin(\varphi - \alpha)|$ depends on α, so the conditional probability depends on the orientation of fiber i as well as on L_i and is no longer a function of the truncated length alone. The invariance is therefore a property of the isotropic, uniformly seeded model, and should not be read beyond it.

2.3. The two models and their first two moments

Throughout, we write

$$c := \frac{2\mathbb{E}(L)}{\pi|C|}, \qquad p := \mathbb{P}(Y_{12} = 1) = c\,\mathbb{E}(L) = \frac{2\mathbb{E}(L)^2}{\pi|C|}, \qquad \mu := \tfrac{1}{2}p, \qquad \sigma^2 := c^2\,\mathrm{Var}(L). \tag{2.5}$$

For the equal-length model (n fibers of common length l), Suh et al. [4] and Chun and Jhong [7, Eq (A.8)] gave

$$m := \mathbb{E}(Y) = n(n-1)\mu, \qquad v := \mathrm{Var}(Y) = \binom{n}{2}p(1-p) + n(n-1)(n-2)\sigma^2, \tag{2.6}$$

together with the central limit theorem $\sqrt{n}\{Y/[n(n-1)] - \mu\} \xrightarrow{d} N(0, \sigma^2)$. For the two-length mixture model (n_g fibers of length l_g, $g = 1, 2$), the corresponding exact mean and variance are recalled in Appendix A. We write $v_{\mathrm{asy}} := n(n-1)^2\sigma^2 = n^3\sigma^2\{1 + O(n^{-1})\}$ for the leading asymptotic variance and reserve v for the exact finite-n variance in (2.6).

Remark 2. *(Terminology) The distinction between v and v_{asy} matters and we keep it visible in the nomenclature. A normal law centered and scaled by (m, v_{asy}) is the* asymptotic Gaussian approximation *delivered by the central limit theorem. A normal law centered and scaled by (m, v) is a* moment-matched Gaussian approximation*: It uses exact finite-n moments and is no longer a purely asymptotic*

object. All comparisons in Section 5 use moment-matched versions of both competitors, so that the two differ only in shape, never in location or scale.

3. Moment-matched negative binomial approximation

3.1. Moment matching

We parametrize $\mathrm{NB}(r, \theta)$, $r > 0$, $\theta \in (0, 1)$, by

$$\mathbb{P}(W = k) = \frac{\Gamma(k + r)}{\Gamma(r)\, k!}\, \theta^r (1 - \theta)^k, \qquad k = 0, 1, 2, \ldots,$$

so that $\mathbb{E}(W) = r(1 - \theta)/\theta$ and $\mathrm{Var}(W) = \mathbb{E}(W)/\theta$. The shape parameter r need not be an integer—the moment-matched values below never are—and the gamma-function form of the coefficient, which reduces to $\binom{k+r-1}{k}$ for integer r, is the one meant throughout.

Lemma 3. *(Moment matching) Let $m > 0$ and $v > m$. The unique $\mathrm{NB}(r, \theta)$ with mean m and variance v has*

$$\theta = \frac{m}{v}, \qquad r = \frac{m^2}{v - m}. \tag{3.1}$$

Proof. From $v = m/\theta$, we get $\theta = m/v$; from $m = r(1-\theta)/\theta$ and $1-\theta = (v-m)/v$, we get $r = m\theta/(1-\theta) = m^2/(v - m)$. Because $v > m > 0$, we have $r > 0$ and $\theta \in (0, 1)$, so the parameters are admissible. Uniqueness holds because (3.1) inverts the injective map $(r, \theta) \mapsto (r(1 - \theta)/\theta,\ r(1 - \theta)/\theta^2)$. □

Lemma 3 applies to any count statistic with known first two moments; it is not specific to the fiber intersection problem. It does require overdispersion, and the exact finite-n threshold is worth recording because the leading-order condition is badly misleading.

The negative binomial is not the only two-parameter lattice family that can be matched to (m, v); the generalized Poisson law of Consul and Jain [9] and the Conway–Maxwell–Poisson law [10, 11] are likewise fixed by two moments, and each implies its own relation between dispersion and skewness. We use the negative binomial for three reasons. It has a closed-form probability mass function and distribution function, whereas the Conway–Maxwell–Poisson normalizing constant must be computed numerically; it is infinitely divisible—a Poisson sum of logarithmic-series variables—which is what makes the cumulant bound of Section 4 available; and it is the classical overdispersed count law, so that the comparison is the one a practitioner would actually make. The choice is not innocuous, because the families differ in the skewness they attach to a given (m, v), and we return to this in Section 5.4 with a numerical comparison.

3.2. Equal-length model and the exact overdispersion threshold

Proposition 1. *(Exact overdispersion threshold) In the equal-length model, $v > m$ if and only if*

$$(n - 2)\sigma^2 > \tfrac{1}{2}p^2, \qquad \textit{and equivalently,} \qquad n > 2 + \frac{p^2}{2\sigma^2} = 2 + \frac{\mathbb{E}(L)^2}{2\,\mathrm{Var}(L)}. \tag{3.2}$$

When it holds, the moment-matched parameters of Lemma 3 are $\theta = m/v$ and $r = m^2/(v - m)$ with m, v from (2.6).

Proof. By (2.6), and using $p = 2\mu$,

$$v - m = \binom{n}{2} p(1-p) - \binom{n}{2} p + n(n-1)(n-2)\sigma^2 = n(n-1)\Big[(n-2)\sigma^2 - \tfrac{1}{2}p^2\Big].$$

The second form follows by substituting (2.5), with the factors of c and $|C|$ cancelling. □

Remark 3. *The threshold* $n > 2 + \mathbb{E}(L)^2/\{2\,\mathrm{Var}(L)\}$ *depends on the geometry only through the dimensionless ratio* $\mathbb{E}(L)^2/\,\mathrm{Var}(L)$*; the counting-region area cancels. For a square counting region of side* 100 *and* $l = 15$*, it reads* $n > 4.08$*, so overdispersion holds for every fiber count of practical interest. By contrast, matching with the* leading-order *variance* $v_{\mathrm{asy}} = n(n-1)^2\sigma^2$ *requires* $\theta_{\mathrm{asy}} = m/v_{\mathrm{asy}} = \mu/\{(n-1)\sigma^2\} < 1$*, i.e.,* $4(n-1)\,\mathrm{Var}(L) > \pi|C|$*; since* $\pi|C|/\{4\,\mathrm{Var}(L)\} = 254.5$ *here, this demands* $n \geq 256$*. The discrepancy is entirely an artifact of discarding the* $O(n^2)$ *Bernoulli diagonal term, which dominates v at moderate n: at* $n = 150$*, that term contributes* 63% *of v. Accordingly, every negative binomial parameter in this paper is obtained from the exact moments* (m, v)*, never from* (m, v_{asy})*.*

Remark 4. *(Below the threshold) When* (3.2) *fails, Y is equidispersed or underdispersed and no negative binomial law has its moments. The regime is not empty: The threshold* $2 + \mathbb{E}(L)^2/\{2\,\mathrm{Var}(L)\}$ *grows as the fibers become short relative to the window, because the truncated length then concentrates at l and* $\mathrm{Var}(L)$ *collapses—for the square of side* 100*, it is* $n > 4.08$ *at* $l = 15$*, but* $n > 8.3$ *at* $l = 5$*,* $n > 17.8$ *at* $l = 2$*, and* $n > 33.7$ *at* $l = 1$*—so a sparse web of short fibers can sit below it. There Y is a sum of weakly dependent indicators of small probability, and the natural reference laws are the Poisson law with mean m when* $v \approx m$ *and, for* $v < m$*, a binomial law with N trials and success probability* m/N*, where N is* $m^2/(m - v)$ *rounded to an integer, which matches the mean exactly and the variance up to the rounding. Poisson approximation for such dissociated sums is the subject of Barbour, Holst, and Janson [12], whose bounds apply to Y directly. Every configuration in Section 5 lies far above the threshold, and the recommendation below is conditional on* (3.2)*.*

3.3. Two-length mixture model

The same principle applies to the mixture model. By Lemma 2, the moment formulas hold under either seeding convention, so a single proposition covers both. For the asymptotic part of the statement, we fix the fiber lengths, the counting region, and the seeding regions, and let the two fiber counts grow in fixed proportion: $n = n_1 + n_2$ with $n_g/n \to \lambda_g \in (0, 1)$, $\lambda_1 + \lambda_2 = 1$. Writing

$$\bar{e} = \lambda_1 \mathbb{E}(L_1) + \lambda_2 \mathbb{E}(L_2), \qquad \bar{v} = \lambda_1\,\mathrm{Var}(L_1) + \lambda_2\,\mathrm{Var}(L_2), \tag{3.3}$$

the mixture analogues of μ and σ^2 in (2.5) are

$$\mu_{\mathrm{mix}} = \frac{\bar{e}^{\,2}}{\pi|C|}, \qquad \sigma^2_{\mathrm{mix}} = \frac{4\,\bar{v}\,\bar{e}^{\,2}}{\pi^2|C|^2} > 0, \tag{3.4}$$

as follows by collecting the leading terms of the mixture mean and variance of Appendix 6: the n^3 coefficient of the variance groups as $\mathrm{Var}(L_1)\lambda_1\bar{e}^{\,2} + \mathrm{Var}(L_2)\lambda_2\bar{e}^{\,2}$.

Proposition 2. *(Moment-matched negative binomial for the two-length mixture) Let $m = \mathbb{E}(Y)$ and $v = \mathrm{Var}(Y)$ be computed from the mixture formulas of Appendix 6, with $\mathbb{E}(L_g)$ and $\mathrm{Var}(L_g)$ evaluated under the seeding region actually used for type g. If $v > m$, the unique negative binomial law W with $\mathbb{E}(W) = m$ and $\mathrm{Var}(W) = v$ has*

$$W \sim \mathrm{NB}(r, \theta), \qquad \theta = \frac{m}{v}, \qquad r = \frac{m^2}{v - m}, \tag{3.5}$$

and, in the fixed-proportion regime $n_g/n \to \lambda_g \in (0, 1)$ of (3.3)–(3.4), with $m \sim n^2\mu_{\mathrm{mix}}$ and $v \sim n^3\sigma^2_{\mathrm{mix}}$,

$$\theta \sim \frac{\mu_{\mathrm{mix}}}{n\,\sigma^2_{\mathrm{mix}}} = O(n^{-1}), \qquad r \sim \frac{n\,\mu^2_{\mathrm{mix}}}{\sigma^2_{\mathrm{mix}}} = O(n) \qquad \text{as } n \to \infty. \tag{3.6}$$

Proof. The first statement is Lemma 3 applied with m, v as given. For the second, $m \sim n^2\mu_{\mathrm{mix}}$ and $v \sim n^3\sigma^2_{\mathrm{mix}}$ give $\theta = m/v \approx \mu_{\mathrm{mix}}/(n\sigma^2_{\mathrm{mix}})$ and $r = m^2/(v - m) \approx n^4\mu^2_{\mathrm{mix}}/(n^3\sigma^2_{\mathrm{mix}}) = n\mu^2_{\mathrm{mix}}/\sigma^2_{\mathrm{mix}}$. □

The same computation in the equal-length model gives $\theta = O(n^{-1})$ and $r \sim n\mu^2/\sigma^2$. In both models, therefore, the shape parameter diverges linearly in the number of fibers while $\theta \to 0$. Section 4 shows what that implies.

Proposition 2, like Lemma 3, is a statement about moments and not about accuracy: It says which negative binomial law is the moment-matched one and how its parameters behave, and nothing about how close that law is to the distribution of Y. We write $Y \approx W$ for the proposal that W be used in place of Y, and treat how well it holds as a question to be settled empirically in Section 5.

4. The Gaussian limit of the negative binomial family

Since the moment-matched shape parameter grows linearly in n, the negative binomial approximation must itself become Gaussian, so that it refines rather than contradicts the central limit theorem of [7]. We now make this precise. Throughout this section, $\kappa_j(X)$ denotes the jth cumulant of a random variable X, which is the coefficient of $(it)^j/j!$ in the cumulant generating function $K_X(t) := \log \mathbb{E}e^{itX}$; the symbol γ_j is reserved for the cumulants of the geometric law in Lemma 4. Under moment matching, r_n and θ_n move together, so what is needed is a limit theorem for jointly varying parameters; the divergence of r_n alone does not suffice.

Example 1. *(Counterexample) Let $W_r \sim \mathrm{NB}(r, \theta_r)$ with $\theta_r = 1 - r^{-2}$. Then $r \to \infty$ but $\mathbb{P}(W_r = 0) = \theta_r^r = (1 - r^{-2})^r \to 1$, so the standardized law degenerates to a point mass and cannot converge to $N(0, 1)$. Consistently,*

$$\mathrm{Skew}(W_r) = \frac{2 - \theta_r}{\sqrt{r(1 - \theta_r)}} \tag{4.1}$$

diverges like $r^{1/2}$. The quantity that must diverge is $r(1 - \theta)$, not r.

Lemma 4. *(Geometric cumulant bound) Let $\gamma_j(\theta)$ denote the jth cumulant of the geometric law on $\{0, 1, 2, \dots\}$ with $\mathbb{P}(G = k) = \theta(1 - \theta)^k$, and write $q = 1 - \theta$. Then for every $j \geq 1$,*

$$\gamma_j(\theta) \leq (j - 1)!\,\frac{q}{\theta^j}, \qquad \text{hence} \qquad \frac{\gamma_j(\theta)}{\gamma_2(\theta)^{j/2}} \leq (j - 1)!\; q^{-(j-2)/2}. \tag{4.2}$$

Proof. For every real $r > 0$, the probability generating function of $\mathrm{NB}(r,\theta)$ is $\{\theta/(1-qs)\}^r = \exp[\, r\{\log\theta - \log(1-qs)\}\,] = \exp[\, r\sum_{k\geq 1} q^k(s^k-1)/k\,]$, so the law is infinitely divisible with Lévy measure $\nu(\{k\}) = r\,q^k/k$, $k \geq 1$; equivalently, $\mathrm{NB}(r,\theta)$ is a Poisson sum of logarithmic-series variables, a representation due to Quenouille [13] that holds for non-integer r as well (see [14, Chapter 5]). Hence $\kappa_j(\mathrm{NB}(r,\theta)) = r\sum_{k\geq 1} k^j\, q^k/k = r\sum_{k\geq 1} k^{j-1}q^k$, and taking $r = 1$ gives $\gamma_j(\theta) = \sum_{k\geq 1} k^{j-1}q^k$. Since $\binom{k+j-2}{j-1} = \prod_{i=0}^{j-2}(k+i)/(j-1)! \geq k^{j-1}/(j-1)!$,

$$\gamma_j(\theta) \leq (j-1)!\sum_{k\geq 1}\binom{k+j-2}{j-1}q^k = (j-1)!\,\frac{q}{(1-q)^j} = (j-1)!\,\frac{q}{\theta^j}.$$

The second bound follows from $\gamma_2(\theta) = q/\theta^2$, so that $\gamma_2(\theta)^{j/2} = q^{j/2}\theta^{-j}$. □

Proposition 3. *(Gaussian limit) Let $W_n \sim \mathrm{NB}(r_n,\theta_n)$, write*

$$Z_n := \frac{W_n - \mathbb{E}(W_n)}{\sqrt{\mathrm{Var}(W_n)}}, \qquad \varrho_n := r_n(1-\theta_n),$$

and suppose $\varrho_n \to \infty$. Then, for every $j \geq 3$,

$$|\kappa_j(Z_n)| \;\leq\; (j-1)!\;\varrho_n^{-(j-2)/2}, \tag{4.3}$$

and consequently $Z_n \xrightarrow{d} N(0,1)$. Conversely, if $Z_n \xrightarrow{d} N(0,1)$, then $\varrho_n \to \infty$; so the hypothesis is necessary as well as sufficient. Moreover, $\varrho_n \to \infty$ implies $r_n \to \infty$, and the converse fails by Example 1.

Proof. By the infinite divisibility established in the proof of Lemma 4, which holds for every real $r_n > 0$, $\kappa_j(W_n) = r_n\gamma_j(\theta_n)$, so

$$\kappa_j(Z_n) = \frac{\kappa_j(W_n)}{\mathrm{Var}(W_n)^{j/2}} = \frac{r_n\gamma_j(\theta_n)}{\{r_n\gamma_2(\theta_n)\}^{j/2}} = \frac{\gamma_j(\theta_n)}{\gamma_2(\theta_n)^{j/2}}\,r_n^{-(j-2)/2} \leq (j-1)!\,\{r_n(1-\theta_n)\}^{-(j-2)/2},$$

by Lemma 4; this is (4.3). The log-characteristic function $t \mapsto \log\mathbb{E}e^{itW_n} = r_n\{\log\theta_n - \log(1-q_ne^{it})\}$, with $q_n = 1-\theta_n$, is analytic in a complex neighborhood of the origin of radius $-\log q_n \geq \theta_n$, so $K_{Z_n}(t) = \log\mathbb{E}e^{itZ_n}$ admits the convergent expansion $-t^2/2 + \sum_{j\geq 3}\kappa_j(Z_n)(it)^j/j!$ for $|t| < \sqrt{\mathrm{Var}(W_n)}\,(-\log q_n)$, and $\sqrt{\mathrm{Var}(W_n)}\,(-\log q_n) \geq (\sqrt{r_nq_n}/\theta_n)\,\theta_n = \varrho_n^{1/2} \to \infty$. For fixed t and large n,

$$\left|K_{Z_n}(t) + \tfrac{t^2}{2}\right| \leq \sum_{j\geq 3}\frac{(j-1)!\,\varrho_n^{-(j-2)/2}|t|^j}{j!} = \sum_{j\geq 3}\frac{|t|^j}{j\,\varrho_n^{(j-2)/2}} \leq \frac{|t|^3}{3\sqrt{\varrho_n}}\cdot\frac{1}{1-|t|/\sqrt{\varrho_n}} \longrightarrow 0.$$

Hence the characteristic function of Z_n converges pointwise to $e^{-t^2/2}$, and Lévy's continuity theorem gives $Z_n \xrightarrow{d} N(0,1)$.

For the converse, note that $W_n \geq 0$ and $\mathbb{E}(W_n)/\sqrt{\mathrm{Var}(W_n)} = \sqrt{\varrho_n}$, so

$$Z_n \;\geq\; -\frac{\mathbb{E}(W_n)}{\sqrt{\mathrm{Var}(W_n)}} \;=\; -\sqrt{\varrho_n} \qquad \text{almost surely.}$$

If $\varrho_n \not\to \infty$, there is a subsequence along which $\varrho_n \leq M < \infty$, and along it, $\mathbb{P}(Z_n < -\sqrt{M}) = 0$ for every n. Since $\Phi(-\sqrt{M}) > 0$, the portmanteau theorem applied to the open set $(-\infty, -\sqrt{M})$ rules out convergence of that subsequence to $N(0,1)$, and hence $Z_n \xrightarrow{d} N(0,1)$ forces $\varrho_n \to \infty$. Finally, $\varrho_n \leq r_n$ because $1-\theta_n \leq 1$. □

Corollary 1. *(The negative binomial approximation nests the Gaussian limit) For the equal-length model with parameters matched to* (m, v)*, we have* $\theta_n = O(n^{-1})$ *and* $r_n \sim n\mu^2/\sigma^2$*, so* $\varrho_n = r_n(1-\theta_n) \sim n\mu^2/\sigma^2 \to \infty$ *and Proposition 3 applies; the same holds for the two-length mixture model with* $n = n_1 + n_2$*, by* (3.6)*. The moment-matched negative binomial approximation therefore reduces to the same Gaussian limit as the central limit theorems of [7], and by* (4.3)*, its own standardized cumulants of order* $j \geq 3$ *depart from the Gaussian values only at order* $\varrho_n^{-(j-2)/2}$*, that is, at order* $n^{-1/2}$ *for* $j = 3$*.*

Two things Corollary 1 does not say should be stated plainly. The rate $O(n^{-1/2})$ is a statement about the cumulants of the approximating law alone; it bounds how far the negative binomial is from normality, not how far it is from the law of Y, and it therefore carries no Kolmogorov or total-variation error bound for the approximation itself. Sharing a Gaussian limit does not determine which of the two candidates is closer to the law of Y at finite n: The cumulant bound is a statement about the negative binomial law alone and fixes neither the sign nor the size of its approximation error relative to Y. That is an empirical question. Section 5 answers it for the configurations studied there; a distributional error bound, which the present tools do not give, is discussed in Section 6.

The negative binomial–normal relation has of course been studied in its own right, and more sharply than we need here: Ouimet [15] gave a local limit theorem with a refined continuity correction for $\mathrm{NB}(r, \theta)$, while Brown and Phillips [16] and Barbour, Gan, and Xia [17] worked in the other direction, bounding the distance from a dependent count to a negative binomial law. Proposition 3 claims no sharpness; it identifies which parameter must diverge once r_n and θ_n are tied together by moment matching, which is the regime the fiber model puts us in. The classical statements of the normal limit for this family (see [14, Chapter 5]) let $r \to \infty$ with θ fixed, which is the sufficiency half under a stronger hypothesis; we have not found the criterion $\varrho_n \to \infty$ stated in the present necessary-and-sufficient form, but we would not claim more for it than that it is the natural formulation for jointly varying parameters. Its content is modest. Since $\varrho_n = \mathbb{E}(W_n)^2/\operatorname{Var}(W_n)$, the condition says that the coefficient of variation of W_n tends to zero; the necessity half then holds for any family of nonnegative laws, by the argument in the proof, and only the sufficiency half uses the negative binomial structure through Lemma 4.

5. Simulation study

5.1. Design

All experiments use the geometric Monte Carlo procedure of Suh et al. [4]: Fibers are generated with midpoints uniform over the relevant seeding region and orientations uniform on $[0, \pi)$, every pair of segments is tested for intersection in closed form, and a crossing is counted when the intersection point lies inside C. No grid discretization is used anywhere. The counting region C is a square of side 100 throughout. For a fiber of length l, the seeding region is the square of side $100 + l$ concentric with C, except in the common-seeding row of Section 5.5, where by construction both fiber types are seeded on the single square of side 130 dictated by the longer of them. In both cases, (2.1) holds: The square of side $100 + l$ is the dilation of C by the square $[-l/2, l/2]^2$, which contains the disc dilation $C \oplus B(0, l/2)$ without being equal to it, and the side-130 square contains that in turn. Only the area $|S|$ of the seeding region enters any of the formulas above, and by Lemma 1, that is the whole of its influence.

Theoretical moments. Every truncated-length moment used below is taken from the closed forms of Lemma 1; none is estimated. This matters more than its triviality suggests. The obvious alternative, a Monte Carlo evaluation of $\mathbb{E}(L)$ and $\mathrm{Var}(L)$ that clips each fiber against C with the Liang–Barsky algorithm [18], carries a standard error of about 5×10^{-5} in relative terms at 10^8 draws. Displacing $\mathbb{E}(L)$ by 10^{-4}, two such standard errors, changes the Kolmogorov distances of Table 1 by up to one-half at the larger fiber counts and reverses the sign of the comparison at $n = 1200$. The reason is the one identified in Section 5.2: Since $m \propto \mathbb{E}(L)^2$, a relative error of 10^{-4} in $\mathbb{E}(L)$ moves the common center of all three approximations by about a fiftieth of a count at $n = 150$ but by more than a count at $n = 1200$, and at the larger fiber counts, the differences under study are themselves of that order. Where a quantity is known exactly, it should be used exactly. When no closed form is at hand, the covariogram integral of Lemma 1 can still be evaluated by one-dimensional quadrature for any convex C; and when $\mathbb{E}(L)$ must genuinely be estimated, the effect is confined to the large-n comparison. Displacing $\mathbb{E}(L)$ by a relative $\pm10^{-4}$, the two standard errors of the preceding paragraph, changes D at $n = 150$ by ±0.0006 and leaves D_{tail} there unchanged to the third decimal, whereas at $n = 1200$, it moves D_{NB} between 0.0039 and 0.0085 and D_{cc} between 0.0061 and 0.0074, reversing their order. The approximations themselves remain usable with estimated moments; what cannot then be resolved is which of the two is closer at the larger fiber counts, which the next subsection shows is not resolved even with exact moments.

The three competing approximations. All three are built from the same finite-n moments (m, v) of (2.6), so that they differ only in shape:

$$\begin{aligned}
\text{Gaussian:}\quad & F_{\mathrm{N}}(y) = \Phi\big((y-m)/\sqrt{v}\big),\\
\text{Continuity-corrected Gaussian:}\quad & F_{\mathrm{cc}}(k) = \Phi\big((k+\tfrac{1}{2}-m)/\sqrt{v}\big),\ k\in\mathbb{Z},\\
\text{Moment-matched negative binomial:}\quad & F_{\mathrm{NB}} = \text{c.d.f. of } \mathrm{NB}(r,\theta),\ \theta = m/v,\ r = m^2/(v-m).
\end{aligned}$$

Comparing F_{N} with F_{NB} alone, as is often done, confounds two quite different effects: the shape of the approximating law, and the mismatch between a continuous law and a lattice-valued statistic. The continuity-corrected version isolates the first and is the fair competitor for a lattice law. We report F_{N} as well, because it is the comparison usually made—note that it is already a moment-matched Gaussian in the sense of Remark 2, not the asymptotic one the central limit theorem delivers—but we base no conclusion on the F_{N} versus F_{NB} comparison.

The Kolmogorov–Smirnov distance on a lattice. Our global criterion is $D = \sup_{y\in\mathbb{R}}|\hat{F}(y) - F_0(y)|$. Because Y is integer-valued, $\hat{F}$ is constant on $[k, k+1)$ for every integer k, so the supremum over $\mathbb{R}$ is attained at an endpoint of such an interval, and

$$D = \begin{cases} \max_{k\in\mathbb{Z}}|\hat{F}(k) - F_0(k)| & \text{if } F_0 \text{ is supported on } \mathbb{Z},\\ \max_{k\in\mathbb{Z}}\max\{|\hat{F}(k) - F_0(k)|,\ |\hat{F}(k) - F_0(k+1)|\} & \text{if } F_0 \text{ is continuous.}\end{cases} \tag{5.1}$$

The second line is the classical order-statistic formula implemented by standard software. The first is not: Applying the classical formula to a lattice F_0, by comparing $\hat{F}(k-)$ against $F_0(k)$, adds a spurious term of the size of the modal probability mass. The resulting quantity is not a distance at all. Drawing R observations from $\mathrm{NB}(r,\theta)$ and comparing them with $\mathrm{NB}(r,\theta)$ *itself*, it stays between 0.033 and 0.043 for the parameters of our $n = 150$ configuration—close to the modal mass $(2\pi v)^{-1/2} = 0.033$—for

every R from 2000 to 160000, instead of decaying like $R^{-1/2}$. Since two of our three candidate laws are lattice laws, and since the discreteness of Y is precisely what is at issue, we state (5.1) explicitly; the supplementary code carries the same warning and a self-consistency check.

The tail-restricted distance. Alongside D, we report the same supremum restricted to the tails. The restriction is defined from the empirical distribution function of the sample in hand: Writing $\hat{F}_R$ for it,

$$\mathcal{T}_R = \{k \in \mathbb{Z} : \hat{F}_R(k) < 0.05 \ \text{ or } \ \hat{F}_R(k) > 0.95\}, \qquad D_{\text{tail}} = \max_{k \in \mathcal{T}_R} \left|\hat{F}_R(k) - F_0(k)\right|, \tag{5.2}$$

for a lattice F_0; for the uncorrected Gaussian, the inner quantity is replaced by the two-endpoint form of (5.1), as in D. Using $\hat{F}_R$ rather than theoretical quantiles keeps the restriction the same for all three candidate laws, so that the comparison is not affected by which of them is used to locate the tails; the price is that $\mathcal{T}_R$ is itself random. In the paired bootstrap below, $\mathcal{T}_R$ is therefore recomputed from each resample's own empirical distribution function, so that the reported standard error is the uncertainty of the statistic actually tabulated. The alternative, freezing $\mathcal{T}_R$ at the full-sample mask, changes the standard error by only a few percent in the equal-length configurations; in the type-specific mixture of Section 5.5, it lowers it by about a quarter, and we report the larger, recomputed value there as elsewhere.

Why not a p-value. We deliberately do not report goodness-of-fit p-values. A p-value answers the question "is the approximation exactly correct?", to which the answer here is known in advance to be no for all three competitors. Its power grows with the number of Monte Carlo replicates, so that at large R, every approximation is rejected, and at small R, none is. This is not a hypothetical concern at the replicate counts customary in this literature: With $R = 600$, D is of order 0.035 even when the reference law is correct, so observed statistics anywhere in the range 0.025–0.07 carry essentially no information about which of the two approximations fits better.

We therefore report D itself, together with two quantities that say how much of it is noise. The first is a Monte Carlo standard error for the paired difference $D_{\text{NB}} - D_{\text{cc}}$, obtained from 400 nonparametric bootstrap resamples of the simulated counts; because the two statistics are computed on the same resample, this is the relevant uncertainty for the comparison, and it is what we base conclusions on. The second is $D_0 = 0.8687/\sqrt{R}$, the asymptotic mean of the Kolmogorov–Smirnov statistic when the reference law is correct and continuous , which we tabulate as a scale marker only. It is not an exact expectation, and it is not a floor: for a lattice reference law, the null distribution of D depends on the jumps of F_0 and has to be computed from them [19], and in that case, D_0 typically overstates the noise level. We use it to answer the coarse question of whether a reported distance is of the size that sampling error alone would produce, and never to decide between two approximations. Replicate counts R are between 12000 and 40000. The sample at $n = 1200$ is the pooling of two independent runs of 8000 and 16000 replicates; the second was added because a single run of 8000 was too small to separate the competitors at that fiber count, and both random seeds are recorded in the supplement. The two runs agree: Their means are 5891.3 and 5892.9 (0.6 standard errors apart) and their variances 33249 and 33246, and the paired difference is -0.0002 ± 0.0024 in the first and $+0.0027 \pm 0.0023$ in the second, both consistent with the pooled $+0.0014 \pm 0.0022$.

Two checks on the inference. Because D is a supremum, two cautions apply to the bootstrap, and we have checked both. The first concerns the standard error. With the fitted negative binomial at $n = 150$ and at $n = 800$ taken as a known lattice truth, we drew 200 independent samples of the same replicate count R and bootstrapped each (200 resamples): The mean bootstrap standard error

of $D_{\mathrm{NB}} - D_{\mathrm{cc}}$ was 1.09 and 0.85 times the Monte Carlo standard deviation of the estimate across samples, and an m-out-of-n bootstrap with $m = R/4$ gave values of the same order. The second concerns bias: A supremum of noisy deviations is biased upward, in that experiment by about 0.003 for the law that is the truth and has distance zero, which is the size of D_0; there, this brought the coverage of nominal 95% intervals for the population difference down to 0.90 at $n = 150$ and 0.70 at $n = 800$, where the true difference is itself of the size of the bias. Subsample trajectories say the same from the data: Evaluating D on random subsamples of $R' = R/16, \ldots, R$ replicates and extrapolating $D(R') = D_\infty + cR'^{-1/2}$ gives $D_\infty = 0.019$ and 0.007 for the corrected Gaussian and the negative binomial at $n = 150$, 0.0045 and 0.000 at $n = 300$, and, from $n = 500$ onward, values between −0.003 and +0.004 for both laws, which is zero to within the precision of the extrapolation. The calibration experiment thus revealed undercoverage of the nominal 95% intervals, and we draw the consequence for everything that follows. Bootstrap standard errors are reported as descriptive measures of Monte Carlo variability, not as the basis for calibrated significance statements; the bias estimate obtained under the negative binomial calibration model is a property of that experiment and is not treated as a bias bound, in either direction, for the fiber-count distribution; and differences of the size of D_0 are treated as unresolved throughout. The reductions in distance reported below are empirical findings of the stated size. The supplementary material contains both checks.

5.2. Global fit: The Kolmogorov–Smirnov distance

Table 1. Kolmogorov–Smirnov distance between the simulated distribution of Y and each approximation (equal-length model, $l = 15$, square C of side 100), computed by (5.1). r and θ are the moment-matched negative binomial parameters of Lemma 3 at the exact finite-n moments. "Gauss." is the uncorrected Gaussian and "Gauss.$_{cc}$" is its continuity-corrected version, both centered and scaled at the same moments. D_{tail} restricts the supremum to the outer 10% of the distribution. The penultimate column is the paired difference on the full sample with a bootstrap standard error (400 resamples); it is the basis for every comparison drawn from this table. The last column, $D_0 = 0.8687/\sqrt{R}$, is the asymptotic mean of D for a correct continuous reference law, given as a scale marker only (see Section 5.1).

				D (whole range)			D_{tail} (outer 10%)				
n	R	r	θ	Gauss.	Gauss.$_{cc}$	NB	Gauss.	Gauss.$_{cc}$	NB	$D_{\mathrm{NB}} - D_{cc}$	D_0
150	40000	159.6	0.6356	0.03718	0.02098	0.01021	0.01042	0.00720	0.00156	−0.01077 (0.00036)	0.00434
300	30000	315.9	0.4624	0.01545	0.00881	0.00446	0.00739	0.00548	0.00210	−0.00436 (0.00257)	0.00502
500	20000	524.3	0.3391	0.01000	0.00779	0.00881	0.00639	0.00538	0.00268	+0.00101 (0.00215)	0.00614
800	12000	837.0	0.2423	0.00753	0.00562	0.00560	0.00308	0.00278	0.00213	−0.00001 (0.00285)	0.00793
1200	24000	1253.8	0.1755	0.00591	0.00485	0.00620	0.00266	0.00238	0.00149	+0.00135 (0.00215)	0.00561

Table 1 reports the central empirical result. Read the $n = 150$ row from left to right. The uncorrected moment-matched Gaussian gives $D = 0.0372$. Adding the continuity correction, and changing nothing else, brings this down to 0.0210. Replacing the corrected Gaussian by the moment-matched negative binomial brings it down again to 0.0102. The two steps are of comparable size, the first being worth a factor of 1.8 and the second a factor of 2.1, and they are logically independent: The first is about support, the second about shape.

Both steps deserve to be taken, and the order in which they are taken decides what the comparison appears to show. A study that compares an uncorrected normal law against a negative binomial attributes the whole of the improvement from 0.0372 to 0.0102 to the change of distributional family, when in fact 60% of it is bought by an adjustment of one half-count that costs nothing and requires no new distribution. That is worth stating plainly because it is the natural comparison to make and it overstates the case for the negative binomial by about a factor of two. But it overstates a case that is nevertheless real at the smaller fiber counts. The paired difference $D_{\mathrm{NB}} - D_{\mathrm{cc}}$ is -0.0108 at $n = 150$, a reduction of the distance from 0.0210 to 0.0102. Its bootstrap standard error, 0.0004, describes the Monte Carlo variability of the difference, and the extrapolation in R of Section 5.1 ($D_\infty = 0.019$ and 0.007) indicates that the reduction is not an artifact of the finite replicate count; for the reasons given there, we report it as an empirical finding of that size and attach no calibrated significance statement to it. At $n = 300$, the difference is -0.0044 ± 0.0026, of the order of D_0, with the extrapolation (0.0045 against 0.000) pointing the same way; we read it as suggestive rather than decisive.

From $n = 500$ onward, that margin disappears. The paired difference is $+0.0010 \pm 0.0022$ at $n = 500$, -0.00001 ± 0.0029 at $n = 800$, and $+0.0014 \pm 0.0022$ at $n = 1200$: All three are within one standard error of zero, and the sign is not stable across them. On the global criterion, therefore, the two lattice approximations are indistinguishable at these fiber counts with these replicate counts, and we draw no conclusion from the ordering of the entries. Two facts reinforce that reading. The reported distances there are of the same size as D_0, so sampling error alone would produce values of this magnitude. In addition, the comparison is sensitive at that scale to sub-count displacement of the center: All three approximations are centered at the theoretical mean m, and the simulated mean sits between 0.25 and 0.9 counts away from it in these rows, a displacement of the same order as the half-count continuity correction itself and one that does not affect the two lattice laws equally, since they place their mass differently about a common center.

The size of that effect can be gauged directly. Recomputing both lattice approximations from the simulated mean and variance instead of the theoretical ones gives D_{NB} against D_{cc} of 0.0056 against 0.0162 at $n = 150$, 0.0038 against 0.0103 at $n = 300$, 0.0030 against 0.0081 at $n = 500$, 0.0046 against 0.0066 at $n = 800$, and 0.0041 against 0.0058 at $n = 1200$: The negative binomial is then ahead at every fiber count. This is a sensitivity analysis and not a correction. Refitting on the same data used to evaluate the fit answers a different question from the one the table answers, and it cannot show that the entries at $n \geq 500$ are "really" in the negative binomial's favor. What it does show is that the ordering of two quantities this close is sensitive to the choice of location and scale, and is therefore not a stable finding of the global criterion. We report the table computed from the theoretical moments, since those are what a user of the approximation would have.

It is useful to place these numbers in the frame the theory of U-statistics provides. Y is a nondegenerate U-statistic of order two, so the Kolmogorov distance between its law and the Gaussian is of order $n^{-1/2}$ by the Berry–Esseen theorem for U-statistics [20, 21], and the leading correction to the Gaussian is the skewness term of the Edgeworth expansion, of the same order [22, 23]. The observed D_{cc} decays like $n^{-0.67}$ over the range of Table 1 ($n^{-0.85}$ between 150 and 500, flattening as the noise floor is reached), consistent with that rate. The same frame says how much room for improvement there is. At $n = 150$, the first Edgeworth correction to the distribution function has modulus at most $|\mathrm{Skew}(Y)|\varphi(0)/6 \approx 0.015$ with the simulated skewness of Section 5.4, and 0.012

with the skewness the negative binomial implies; the corrected Gaussian is 0.021 from the simulated distribution, and the negative binomial recovers 0.011 of that. The achievable gain is thus of the order of the skewness term, and the negative binomial realizes most of it at the fiber counts where that term is large.

Figure 1 displays the two lattice approximations directly, and shows what the numbers above describe. At $n = 150$, the negative binomial curve stays markedly near to zero across the whole range, and the two are never confusable. At $n = 300$, the separation is still visible, but the curves have begun to overlap. At $n = 800$, they are interleaved, of comparable amplitude, and neither is near to zero than the other except over short stretches—which is the picture the paired difference reports as indistinguishable. Note also that the vertical scale falls by roughly a factor of four across the three panels: What is disappearing is not the ordering alone but the size of the discrepancy being ordered.

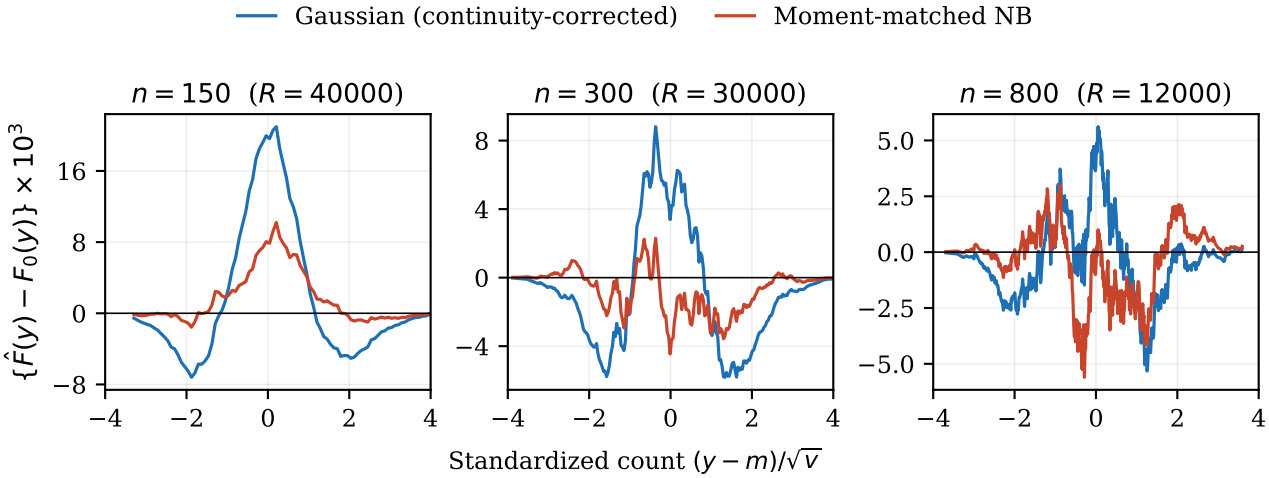


Figure 1. Difference between the simulated distribution function of Y and each approximation, against the standardized count. Both approximations are built from the same finite-n moments (m, v) and so differ only in shape.

5.3. *Tail accuracy*

The tails are where an approximation to a count statistic is actually used: Confidence intervals for $\mathbb{E}(Y)$, prediction intervals for Y itself, and critical values for tests of a hypothesized fiber density are all tail quantities. There, the absolute differences of Section 5.2 become small, but the probabilities they are differences of become smaller still.

Restricting the Kolmogorov supremum to the outer 10% of the distribution already shows this. In Table 1, the column D_{tail} favors the negative binomial at every one of the five fiber counts, and unlike the unrestricted D, it does not change sign anywhere: The ratio $D_{\text{tail,NB}}/D_{\text{tail,cc}}$ is 0.22, 0.38, 0.50, 0.77, and 0.63 as n runs from 150 to 1200. The same paired bootstrap, applied to D_{tail}, puts the difference at 7.8, 13.8, and 14.3 standard errors for $n = 150, 300, 500$. At $n = 800$ and $n = 1200$, the difference keeps its sign but is within one standard error (0.5 and 0.8), so those two rows are consistent with a smaller version of the same effect and also with none; we do not claim more from them. The contrast with the global column is the point: Where the unrestricted D has the two approximations changing places from $n = 500$ onward, the restricted one keeps the negative binomial ahead throughout.

The outer 10% is a conventional choice, the 5% and 95% points being where tail quantities are usually read; tightening the mask to the outer 2% and 1% (the cells where $\widehat{F}_R$ is below 0.01 or above 0.99, and below 0.005 or above 0.995) leaves the sign of the paired difference unchanged up to $n = 800$—it is -0.0033 ± 0.0006 and -0.0025 ± 0.0005 at $n = 150$, and -0.0015 ± 0.0007 and -0.0002 ± 0.0006 at $n = 800$, the last being too small to be read either way—but $n = 1200$ reverses it, by $+0.0008 \pm 0.0004$ and $+0.0007 \pm 0.0003$, so the advantage in the extreme tails at the two largest fiber counts is not established. Two further qualifications belong here. The statement that the negative binomial is ahead "at every fiber count" concerns the supremum D_{tail} over the two-sided mask and not every cell of Table 3. Up to $n = 500$, the negative binomial is the closer law in all six cells of each row, and at $n = 800$, it is in five (in the upper $z = 2.326$ cell, both laws are within one standard error of the simulated value). In the $n = 1200$ row, the corrected Gaussian is closer in three cells, most clearly in the lower tail at $z = 2.326$ (0.01001 against a simulated 0.00996, with the negative binomial at 0.00890), and also in the upper tail at $z = 2.326$ and $z = 3.0$, where the negative binomial overshoots the simulated probability. The Monte Carlo standard errors of the extreme quantiles in Table 2 grow with the spread of Y, from half a count at $n = 150$ to 4.5 counts at $n = 1200$, so the larger quantile errors in the last row are partly a statement about resolution.

Table 3 shows what that restriction is picking up. For $n \leq 500$, the pattern is systematic and is the signature of a positive third cumulant: the true distribution has a heavier upper tail and a lighter lower tail than a symmetric law of the same mean and variance, so the continuity-corrected Gaussian understates the upper tail and overstates the lower one, and does so by a factor rather than by a small additive amount as one moves outward. Across the eighteen tail cells with $n \leq 500$ in Table 3, the negative binomial is the closer of the two in every one, typically by a factor of three to ten. At $n = 150$ and $z = 3$, the Gaussian understates the upper tail by a factor of 2.2 (6 Monte Carlo standard errors), while the negative binomial errs there by 15%. The corresponding lower-tail cell is the most extreme in the table—the Gaussian overstates it by a factor of 6.0, the negative binomial by a factor of 2.2—but it rests on nine simulated exceedances out of 40000, and multipliers computed from nine events should be read as indicating a direction and an order of magnitude rather than a value. The directional statement is the robust one, and it holds across the whole $n \leq 500$ block. Figure 2 shows the upper-tail relative error as a continuous function of the cut-off.

Table 2. Signed error, in counts, of the approximate 1%, 5%, 95%, and 99% quantiles relative to the simulated quantiles (equal-length model, $l = 15$, square C of side 100), with bootstrap Monte Carlo standard errors of the simulated quantiles.

	Gaussian$_{cc}$ − simulated				NB − simulated			
n	1%	5%	95%	99%	1%	5%	95%	99%
150	−1	+0	−1	−3	+0	+0	+0	−1
(s.e.)	(0.4)	(0.4)	(0.2)	(0.5)	(0.4)	(0.4)	(0.2)	(0.5)
300	−1	−1	−1	−3	+1	+0	+0	−1
(s.e.)	(0.7)	(0.5)	(0.5)	(0.7)	(0.7)	(0.5)	(0.5)	(0.7)
500	−5	−3	−3	−4	−1	−1	−2	+0
(s.e.)	(1.4)	(0.9)	(0.9)	(1.7)	(1.4)	(0.9)	(0.9)	(1.7)
800	−8	−2	−2	−3	−3	+1	+0	+3
(s.e.)	(2.9)	(2.0)	(2.1)	(3.6)	(2.9)	(2.0)	(2.1)	(3.6)
1200	−1	−4	−3	+2	+7	−1	+0	+10
(s.e.)	(4.5)	(2.7)	(2.0)	(4.0)	(4.5)	(2.7)	(2.0)	(4.0)

Table 3. Tail probabilities of the simulated distribution of Y and of the two lattice approximations, at cut-offs $m \pm z\sqrt{v}$ (equal-length model, $l = 15$, square C of side 100). Binomial Monte Carlo standard errors of the simulated values are in parentheses.

		upper tail $\mathbb{P}(Y \geq m + z\sqrt{v})$			lower tail $\mathbb{P}(Y \leq m - z\sqrt{v})$		
n	z	simulated (s.e.)	Gauss.$_{cc}$	NB	simulated (s.e.)	Gauss.$_{cc}$	NB
150	1.645	0.05185 (0.00111)	0.04796	0.05276	0.04190 (0.00100)	0.04762	0.04182
	2.326	0.01405 (0.00059)	0.00986	0.01331	0.00595 (0.00038)	0.00977	0.00633
	3.0	0.00302 (0.00027)	0.00136	0.00258	0.00022 (0.00007)	0.00134	0.00049
300	1.645	0.05643 (0.00133)	0.05063	0.05388	0.04313 (0.00117)	0.04837	0.04469
	2.326	0.01327 (0.00066)	0.01036	0.01271	0.00847 (0.00053)	0.00977	0.00750
	3.0	0.00210 (0.00026)	0.00141	0.00220	0.00073 (0.00016)	0.00131	0.00071
500	1.645	0.05640 (0.00163)	0.05086	0.05335	0.04520 (0.00147)	0.05020	0.04747
	2.326	0.01160 (0.00076)	0.00993	0.01167	0.00775 (0.00062)	0.01025	0.00849
	3.0	0.00180 (0.00030)	0.00134	0.00191	0.00095 (0.00022)	0.00139	0.00090
800	1.645	0.05167 (0.00202)	0.04991	0.05189	0.04867 (0.00196)	0.05001	0.04790
	2.326	0.01058 (0.00093)	0.00993	0.01130	0.00783 (0.00080)	0.00996	0.00860
	3.0	0.00158 (0.00036)	0.00134	0.00178	0.00100 (0.00029)	0.00134	0.00096
1200	1.645	0.05162 (0.00143)	0.04979	0.05140	0.04813 (0.00138)	0.05011	0.04840
	2.326	0.00975 (0.00063)	0.01007	0.01119	0.00996 (0.00064)	0.01001	0.00890
	3.0	0.00100 (0.00020)	0.00135	0.00170	0.00112 (0.00022)	0.00134	0.00102

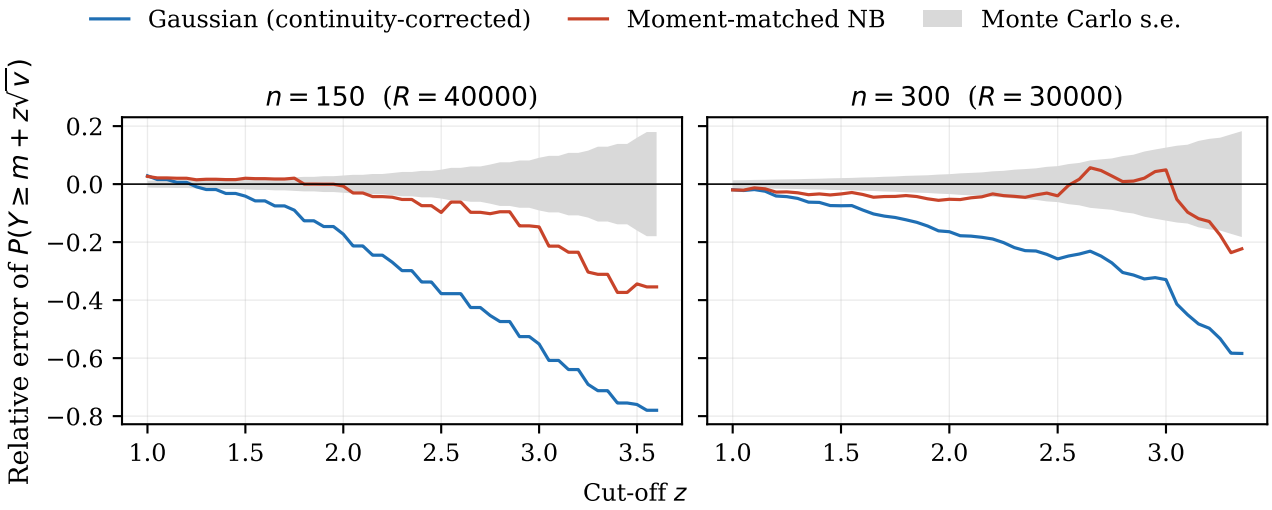


Figure 2. Relative error of the approximate upper-tail probability $\mathbb{P}(Y \geq m + z\sqrt{v})$ as a function of the cut-off z, for two fiber counts, shown only as far out as the simulation resolves (at least 25 *observed* exceedances in the simulation). The shaded band is one estimated relative Monte Carlo standard error of the simulated tail probability, on either side of zero; it is not a confidence band for either approximation curve. The symmetric Gaussian approximation understates the upper tail increasingly severely as z grows. The negative binomial reduces that underestimation over the whole of the displayed range, by a factor of roughly two at the largest cut-offs, but does not remove it: Its own relative error is negative and growing at the edge of resolution.

The same appears in the quantiles, which is the form in which the difference reaches a practitioner. At $n = 150$ and $n = 300$, the negative binomial reproduces the simulated 1% and 99% quantiles to within one count, against Monte Carlo standard errors of half a count at $n = 150$ and seven-tenths of a count at $n = 300$, whereas the continuity-corrected Gaussian is short by three counts at the 99% point in both rows—roughly six and 4.5 standard errors, and in the direction a positive third cumulant predicts. At $n = 500$, the Gaussian errors are between 2.4 and 3.6 standard errors, and the negative binomial's are between 0 and 2.2. At $n = 1200$, the two change places: the negative binomial is out by +7 and +10 counts at the 1% and 99% points (1.6 and 2.5 standard errors) against the Gaussian's −1 and +2 (both under one). This is the overshoot examined in Section 5.4, and it is the only row of Table 2 in which the negative binomial is the worse of the two by more than a Monte Carlo standard error.

5.4. Where the difference comes from

The two lattice approximations are constructed from the same target finite-sample moments (m, v) and, by Corollary 1, share a Gaussian limit. They differ in their third-order behavior, and there the negative binomial has no freedom left: Once m and v are fixed, *every* negative binomial law with those moments has the third cumulant $2v^2/m - v$, hence it has the skewness $(2 - \theta)\{r(1 - \theta)\}^{-1/2}$, while the normal law $N(m, v)$ from which F_{cc} is built has a third cumulant of zero. (Rounding need not preserve these moments and cumulants exactly: We compare against the target (m, v) and the normal value 0 throughout, and make no claim that F_{cc} reproduces them identically.) The question is therefore

not which family is more appropriate for a count, but which of two specific numbers is closer to the skewness of Y.

Simulation answers it directly. At $n = 150$, the simulated skewness is 0.2316 against 0.1789 implied by the fitted NB(159.6, 0.6356); at $n = 300$, it is 0.1263 against 0.1180; at $n = 500$, it is 0.0925 against 0.0892; at $n = 800$, it is 0.0467 against 0.0698; and at $n = 1200$, it is 0.0149 against 0.0567. The Monte Carlo standard errors, 0.0122, 0.0141, 0.0173, 0.0224, and 0.0158, are the normal-theory values $\sqrt{6/R}$; since they are derived under symmetry, they are indicative only, and are quoted to fix the scale rather than to test anything.

Two things are visible. First, in the four rows up to $n = 800$, the negative binomial value is nearer to the simulated one than zero is—at $n = 800$, it overstates the skewness by 0.0231 where the Gaussian understates it by 0.0467. The skewness comparison offers a possible explanation for part of the discrepancy seen in Sections 5.2 and 5.3, but it does not by itself determine the ordering of the approximations in Kolmogorov distance or in tail error. Second, the sign of the discrepancy turns over between $n = 500$ and $n = 800$: The negative binomial moves from understating the skewness of Y to overstating it. The same turnover is visible in the far upper tail of Table 3, where at $z = 3$, the negative binomial falls short of the simulated probability at $n = 150$ and exceeds it at every larger fiber count. At $n = 1200$, the fitted negative binomial has greater skewness than the sample estimate, and it also has the larger observed error at the 99% quantile; the present analysis does not establish a causal relation between these two discrepancies.

It is worth being explicit about what this does and does not establish. The estimated skewness difference changes sign between the $n = 500$ and $n = 800$ configurations. The Monte Carlo uncertainty is far too large to locate a crossover in the population skewness: The simulated skewness minus the negative binomial one is $+0.0033$ at $n = 500$ and -0.0231 at $n = 800$, against standard errors of 0.017 and 0.022, and recomputing those standard errors from the empirical influence function rather than the normal-theory formula changes them by less than 20%. We therefore read the sign change as a description of the simulated configurations, not as an estimate of a turning point. Both skewnesses shrink as the fiber count grows, and the simulated column shrinks with them: At $n = 800$ and $n = 1200$, the simulated skewness is itself only 2.1 and 0.9 standard errors from zero, so the experiment can no longer say with confidence that Y is skewed at all, let alone by how much. How the skewness of Y behaves in n is a question about the third cumulant of a U-statistic with a Bernoulli kernel, which we do not analyze here. The structure of that question can nevertheless be stated. The negative binomial's third cumulant $2v^2/m - v$ grows like n^4, with a leading coefficient $2\sigma^4/\mu$ fixed by the second-order constants of (2.5); the third cumulant of Y also grows like n^4, since the Hoeffding projection of a U-statistic of order two contributes at that order, but its leading coefficient is a different geometric constant, not determined by $\mathbb{E}(L)$ and $\mathrm{Var}(L)$. Whether the moment-matched law overstates or understates the skewness of Y for large n is therefore a fixed property of the geometry, and the terms of lower order in n can reverse the sign at moderate n—which is the pattern the simulated skewness displays, an understatement up to $n = 500$ turning into an overstatement beyond. Settling the sign and the crossover analytically requires the constant itself, and we return to what that entails in Section 6.

The Edgeworth expansion as a third candidate. Since the two lattice laws differ precisely in their third cumulant, the natural benchmark is the lattice Edgeworth approximation carrying the skewness of Y itself, $F_{\mathrm{E}}(k) = \Phi(z_k) - \frac{1}{6}\, g\, (z_k^2 - 1)\varphi(z_k)$ with $z_k = (k + \frac{1}{2} - m)/\sqrt{v}$, whose first term is the continuity-corrected Gaussian. The skewness of Y is not available in closed form, so the simulated skewness

stands in for it below; it is an estimate carrying the Monte Carlo standard errors quoted above, not the true value. With g set to the skewness the negative binomial implies, this expansion reproduces D_{NB} to four decimals at every fiber count in Table 1: To this order, the moment-matched negative binomial *is* the Edgeworth correction with the skewness forced to $(2-\theta)\{r(1-\theta)\}^{-1/2}$. With g set to the simulated skewness—a plug-in that uses the sample it is evaluated on, so it is a sensitivity analysis rather than a usable approximation—the expansion gives $D = 0.0070$ at $n = 150$ against 0.0102 for the negative binomial, 0.0050 against 0.0045 at $n = 300$, and values within one standard error of both competitors from $n = 500$ on. The gain at $n = 150$ is consistent with the understated skewness of the negative binomial there, though it cannot be attributed to that alone, since the plug-in is fitted on the sample it is scored on, and it comes with the usual defect of Edgeworth series: at $n = 150$ and $n = 300$, the expansion is not a distribution function (its lower-tail mass at three standard deviations is -0.00002 at $n = 150$), and in the outer 10%, it is worse than the negative binomial at $n = 150$ ($D_{\mathrm{tail}} = 0.0029$ against 0.0016). The negative binomial is a proper lattice law whose implied skewness is close enough to the simulated value that the expansion with that value improves on it only at the smallest fiber count, and then not in the tails.

Other two-parameter lattice families. The generalized Poisson law matched to (m, v) implies a skewness very close to the negative binomial's (0.184 against 0.179 at $n = 150$) and is empirically indistinguishable from it, with $D = 0.0099$, 0.0050, 0.0092, 0.0062, and 0.0067 across the five fiber counts. The Conway–Maxwell–Poisson law implies a markedly smaller skewness (0.13 at $n = 150$, 0.03 at $n = 1200$): It is worse at $n = 150$ ($D = 0.0131$) and in the outer 10% at every $n \leq 500$ (0.0031, 0.0032, and 0.0038 against 0.0016, 0.0021, and 0.0027), level with it on the global criterion at $n = 300$ (0.0045 against 0.0045), and marginally better there from $n = 500$ on, where the negative binomial's skewness begins to overshoot. The choice between the families is thus the choice of a skewness–dispersion relation, and in the regime where the approximations differ materially, the negative binomial's is the better one.

5.5. Two-length mixture model and seeding-region invariance

We now turn to the mixture model, in the baseline configuration of Chun and Suh [6] and Chun and Jhong [7]: $n_1 = 600$ fibers of length $l_1 = 10$ and $n_2 = 400$ fibers of length $l_2 = 30$ in a square counting region of side 100. Table 4 reports this configuration under both seeding conventions.

Table 4. Two-length mixture model ($n_1 = 600$, $l_1 = 10$; $n_2 = 400$, $l_2 = 30$; square C of side 100) under the common seeding convention of [7] and under type-specific seeding. In both cases, the theoretical moments are computed from the *same* formulas (Appendix 6) with $\mathbb{E}(L_g), \mathrm{Var}(L_g)$ evaluated under the seeding region actually used, as Lemma 2 permits. The close agreement with the simulation in both rows is the numerical confirmation of that lemma.

Seeding	R	$\mathbb{E}(L_1)$	$\mathbb{E}(L_2)$	m (theory)	$\hat{m}$ (sim.)	v (theory)	$\hat{v}$ (sim.)	(r, θ)
Common S [7]	15000	5.9172	17.7515	3606.3	3609.1	37383	37634	(385.0, 0.0965)
Type-specific S_1, S_2	15000	8.2645	17.7515	4623.7	4625.1	44132	44566	(541.1, 0.1048)

The two conventions give genuinely different probability models—the mean intersection count differs by roughly a quarter—but the same moment formulas describe both, with no additional

geometric factor, exactly as Lemma 2 asserts and contrary to the conjecture recorded in [7, Remark 3.2]. This settles the point and, equally usefully, removes a trap: A study that adopts type-specific seeding in its simulations while quoting mean and variance computed from commonly seeded truncated-length moments will find a discrepancy of tens of percent and may misattribute it to the approximation under study.

Table 5 and Figure 3 repeat the three-way comparison for the mixture model. In the tails, the conclusion is the same as in the equal-length model, and stronger: D_{tail} favors the negative binomial by 41% under type-specific seeding and by 55% under common seeding, and the 99% quantile is short by 4 and 1 counts, respectively, where the continuity-corrected Gaussian is short by 18 and 16. That the effect survives in a heterogeneous web, where the two fiber types contribute very different truncated-length variances, indicates that the mechanism identified for the equal-length model operates here too.

Globally, the two rows say different things, and the difference is instructive rather than troubling. Under type-specific seeding, the two are level: $D_{\text{NB}} = 0.0060$ against $D_{\text{cc}} = 0.0059$, a paired difference of +0.00007 against a bootstrap standard error of 0.0031, and both below the continuous-reference marker $D_0 = 0.0071$ for this replicate count. Under common seeding, the negative binomial is clearly worse, 0.0124 against 0.0056, a difference of +0.0067 against a standard error of 0.0027. The tail column runs the other way in both rows and by a margin several times its own standard error: -0.0024 ± 0.0006 and -0.0029 ± 0.0008. At $n_1 + n_2 = 1000$ fibers, this is the regime already met at $n \geq 500$ in the equal-length model: the approximation error under study is a few thousandths, and the global statistic is sensitive to where the center is placed. Here, that sensitivity is unusually visible. The simulated mean exceeds the theoretical m by 1.4 counts under type-specific seeding and by 2.8 counts under common seeding (0.8 and 1.8 Monte Carlo standard errors), and because the negative binomial is right-skewed, its median sits about three counts below its mean, so under common seeding, the two displacements compound instead of cancelling in a distribution whose standard deviation is 193. The skewness the negative binomial implies under common seeding, 0.102, agrees with the simulated 0.092 ± 0.021, so the third moment at least does not account for the discrepancy. Recomputing both approximations from the simulated moments gives $D_{\text{NB}} = 0.0067$ against $D_{\text{cc}} = 0.0069$ there, and 0.0035 against 0.0078 under type-specific seeding. As in Section 5.2, we report the theoretical-moment version, and we read the recomputation as showing that the global comparison in this configuration is sensitive to the choice of location and scale—not as identifying how much of the original error was a location effect, and not as showing that the negative binomial is really ahead. The tail columns, which are what the two laws are being compared for, are not affected by the ambiguity and favor the negative binomial in both rows.

Table 5. Three-way comparison for the two-length mixture model, under both seeding conventions ($n_1 = 600$, $l_1 = 10$; $n_2 = 400$, $l_2 = 30$; square C of side 100). Each paired difference carries a bootstrap standard error from 400 resamples, computed as in Table 1; for D_{tail}, the mask (5.2) is recomputed on each resample.

		D (whole range)				D_{tail} (outer 10%)			
seeding	R	Gauss.	Gauss.$_{cc}$	NB	$D_{\text{NB}} - D_{cc}$	Gauss.$_{cc}$	NB	$D_{\text{NB}} - D_{cc}$	99% q. err. (cc / NB)
common	15000	0.00664	0.00561	0.01235	+0.00674 (0.00268)	0.00521	0.00236	−0.00285 (0.00080)	−16 / −1
type-specific	15000	0.00683	0.00592	0.00599	+0.00007 (0.00311)	0.00590	0.00348	−0.00241 (0.00064)	−18 / −4

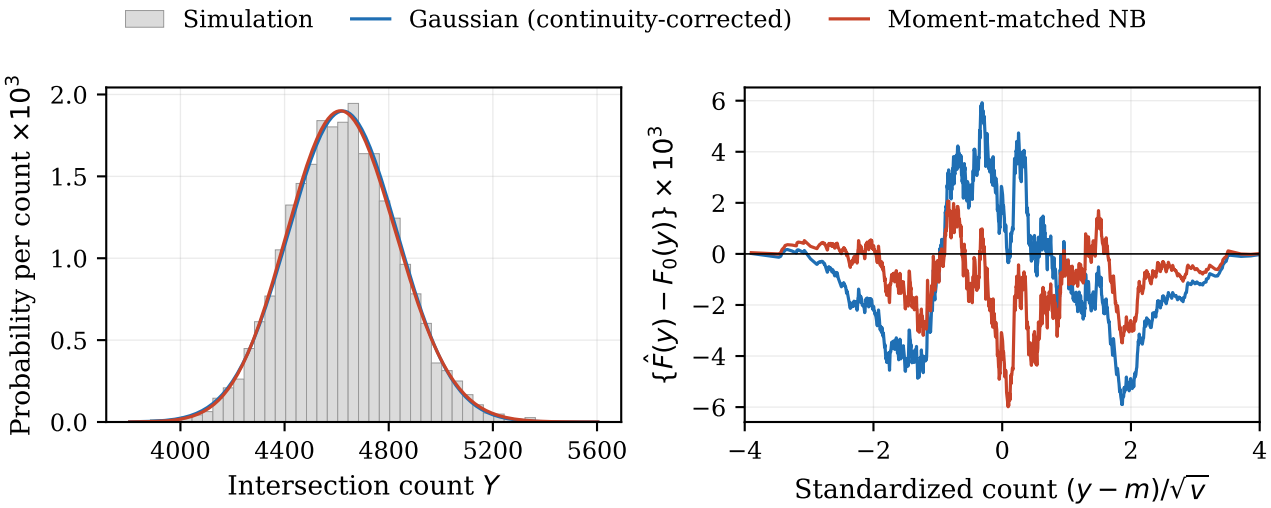


Figure 3. Two-length mixture model under type-specific seeding. Left: The simulated distribution of Y with the two lattice approximations overlaid. Right: The difference between the simulated distribution function and each approximation, as in Figure 1.

5.6. Other counting regions

The square counting region is a convenience, not a restriction. By the remark following Lemma 1, the truncated-length moments of any convex C follow from its isotropic covariogram, and the moment formulas and Lemma 2 do not depend on the shape of C at all. To see whether the comparison does, Table 6 repeats it for a disc and for a 50×200 rectangle of the same area 10^4, with $l = 15$, each seeded on the smallest axis-parallel box containing $C \oplus B(0, l/2)$, at $n = 150$ and $n = 500$. The square is re-run alongside with the same replicate counts so that the three rows are on an equal footing. The shape enters through $\mathbb{E}(L)$ and $\mathrm{Var}(L)$ only. The rectangle, with more boundary per unit area than the square, has a smaller mean and a larger variance of the truncated length; the disc, seeded on its bounding box, has a smaller mean still and a much larger variance, because fibers seeded in the corners of the box never reach it, and its count is correspondingly the most skewed of the three. In every configuration, the picture is the one found for the square. At $n = 150$, the negative binomial has the smaller global distance by 5, 3, and 8 standard errors of the paired difference, and the smaller tail distance by a factor of roughly three to five; at $n = 500$, the global distances are within one or two standard errors of each other, the rectangle's sign reversed exactly as the square's is in Table 1 from $n = 500$ on, while the tail distance still favors the negative binomial for all three regions. At $n = 150$, the simulated skewness exceeds the negative binomial's for all three regions, by 2.7 and 2.8 Monte Carlo standard errors for the square and the rectangle, and by 0.8 for the disc; at $n = 500$, the three differences are within one standard error of zero, as for the square in Section 5.4. The simulated means and variances agree with the covariogram-based moments to within Monte Carlo error.

Table 6. Three counting regions of area 10^4 in an equal-length model with $l = 15$. Truncated-length moments are taken from the covariogram form of Lemma 1, and D and D_{tail} are as in Table 1, with the bootstrap standard error of the paired difference in parentheses. The square rows are a fresh sample at these replicate counts and therefore differ from Table 1 by sampling variation.

							D (whole range)			D_{tail}		Skewness	
Region C	$\mathbb{E}(L)$	$\text{Var}(L)$	n	R	m	v	Gauss.$_{cc}$	NB	s.e.	Gauss.$_{cc}$	NB	sim.	NB
Square 100×100	11.342	30.86	150	20000	91.5	144	0.0163	0.0071	(0.0018)	0.0079	0.0021	0.226	0.179
			500	10000	1021.7	3012	0.0091	0.0052	(0.0026)	0.0067	0.0045	0.100	0.089
Disc, radius 56.42	9.179	45.67	150	20000	59.9	111	0.0159	0.0039	(0.0045)	0.0110	0.0024	0.271	0.257
			500	10000	669.1	2603	0.0098	0.0064	(0.0037)	0.0051	0.0021	0.112	0.133
Rectangle 50×200	10.733	33.17	150	20000	82.0	133	0.0202	0.0081	(0.0014)	0.0097	0.0033	0.243	0.194
			500	10000	915.0	2833	0.0062	0.0077	(0.0038)	0.0047	0.0040	0.096	0.098

6. Conclusions

The number of fiber intersections in a two-dimensional random web is a lattice-valued count, overdispersed in the configurations studied here, and it is tempting to conclude from those two facts alone that a negative binomial law matched to its first two moments must be a better finite-sample description than the Gaussian limit. What the results above show is narrower and, we think, more useful: The conclusion holds in a definite regime, fails to be resolvable outside it, and, in one place, reverses.

Three points are worth carrying away. First, discreteness and shape must be separated before any comparison is meaningful: A negative binomial law lives on the lattice on which Y lives and an uncorrected normal law does not, so comparing the two mixes a question about support with a question about shape. In our experiments, the half-count continuity correction alone accounted for 60% of the apparent improvement from switching families at $n = 150$ and at $n = 300$, and for all of it at the larger fiber counts, where switching families buys nothing further on the global criterion and at $n = 1200$ costs something. Second, the goodness-of-fit statistic itself must be computed for the right kind of reference law; the classical order-statistic form of the Kolmogorov–Smirnov distance assumes continuity, and applied unchanged to a lattice reference, it penalizes exactly the candidate it is meant to assess. Third, once both corrections are made, the advantage of the negative binomial is real but regime-dependent. On the global Kolmogorov criterion, it is worth a factor of two at $n = 150$—a reduction from 0.021 to 0.010 that the subsample extrapolation reproduces—and the same factor at $n = 300$ on evidence that is suggestive only; from $n = 500$ onward, the paired difference between the two fits is within one standard error of zero with an unstable sign. We report those entries but draw no ordering from them. On the Kolmogorov distance restricted to the outer 10% of the distribution, the negative binomial is ahead at every fiber count we ran, by several thousandths up to $n = 500$ and by less than D_0 beyond, though not in every individual tail cell and not, at $n = 1200$, when the mask is tightened to the outer 2% or 1%. In the tail probabilities and extreme quantiles that carry that information, it reproduces the 1% and 99% points to within a count at $n = 150$ and $n = 300$, where the corrected Gaussian is short by three counts at the 99% point. Against that, the negative binomial has the larger 99% quantile error at $n = 1200$, and the larger global distance in the common-seeding mixture configuration. The honest

summary is not that one approximation dominates, but that the two differ where the count is small and the question is a tail question. At the larger fiber counts, our simulations do not resolve a preference on the global criterion; that is a statement about what these experiments can distinguish, and it should not be read as equivalence for every tail probability or quantile of practical interest. At $n = 1200$ and three standard deviations, for instance, the two approximations still put the upper-tail probability at 0.00135 and 0.00170 against a simulated 0.00100—a small difference on the scale of the distribution, but a large one on the scale of the quantity being approximated.

The practical recommendation follows the same division. For central probabilities, a normal law carrying the exact finite-n moments and the half-count correction is adequate at every fiber count we examined, and is analytically far more convenient. For tail probabilities, extreme quantiles, and the confidence and prediction intervals built from them, the moment-matched negative binomial is the better choice at the moderate fiber counts where the two differ materially—in our experiments, up to roughly $n = 500$. Beyond that, the evidence does not separate them, and we would not argue for either on the strength of it.

Two directions are specific to the approximation studied here, and both concern what simulation cannot settle. The size and sign of the residual margin are governed by the third cumulant of Y: The moment-matched negative binomial has no freedom in its third cumulant once m and v are fixed, so whether it overshoots or undershoots, and from which fiber count, is a property of $\kappa_3(Y)$. It is worth saying what its computation involves. The third central moment of the sum of pairwise indicators expands over triples of pairs, classified by how many fibers the three pairs share—the enumeration behind (2.6) and Appendix 6, carried one layer deeper. Beyond $\mathbb{E}(L^3)$, which Lemma 1 supplies, two joint quantities appear that have no counterpart in the second-order theory: An integral coupling the truncated lengths of two fibers to the event that they meet, $\mathbb{E}[(L_1 - \mathbb{E}L)(L_2 - \mathbb{E}L)Y_{12}]$, and the probability that three fibers form a triangle of mutual intersections. The covariogram argument of Lemma 1 handles translates of C along a single fiber and does not extend to the relative position of two or three fibers, so we know of no closed form for either, although both are easily evaluated by Monte Carlo. The task is laborious rather than conceptually difficult, and we are pursuing it separately. It would sharpen the geometry-dependent skewness discrepancy of Section 5.4 into something computable rather than simulated—though a third-cumulant comparison alone would still not establish an ordering in distributional distance or tail-probability error. Separately, an explicit error bound in total variation or Wasserstein distance, in the spirit of the negative binomial Stein method of Brown and Phillips [16] and Barbour et al., [17], and of the refined normal comparisons for the negative binomial family in Ouimet [15], would put the comparison on a footing that no amount of simulation can supply. The Stein factors of [17] are available, and the local dependence of $\{Y_{ij}\}$ is of the kind that method handles; the difficulty is quantitative. Each pair shares a fiber with $2(n - 2)$ others, so the dependency graph has a degree of order n, and bounds that scale with that degree are far larger than the distances of a few thousandths at which the two candidates differ here. An informative bound would have to exploit the Hoeffding structure of Y inside the Stein argument, and we leave it as an open problem. More broadly, the geometric model itself invites extension beyond the two-length mixture to general fiber-length distributions, to curved fibers, and to three-dimensional fiber systems, in each of which the questions raised here would have to be asked again. As such approximations come to be applied to large observation windows and to the massive image data of materials quality control, the computational side of the comparison will matter as well: the Monte Carlo budget here is modest, but at that scale,

the distributed and subsampling-based inference developed for massive survival data [24] suggests how a simulation design of the present kind could be made to preserve its accuracy under a fixed computational budget.

Author contributions

Heuiju Chun: Conceptualization, methodology, formal analysis, investigation, writing–original draft preparation; Jae-Hwan Jhong: Methodology, formal analysis, investigation, software, numerical simulation, writing–review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

During the preparation of this manuscript, the authors used Google Gemini solely for the purpose of English language editing, proofreading, and improving readability. The tool was not employed to generate original text, conceive research ideas, analyze data, or interpret findings. The authors have thoroughly reviewed and edited the manuscript and take full responsibility for the content and scholarly integrity of the publication.

Acknowledgments

This research was supported by the Global-Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2024-00445180). The work of Jae-Hwan Jhong was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00342014).

Data and code availability

All reported numbers were produced by the Python implementation supplied as supplementary material, with the random seeds recorded in each script. The tables and figures were rebuilt from the shipped counts with Python 3.13, NumPy 2.5, SciPy 1.18, and Matplotlib 3.11. The simulated intersection counts are included, and two commands rebuild every table and figure from them without re-simulating anything, one for Tables 1–5 and the figures, and one for Table 6. A third script produces the numbers quoted in Sections 5.1–5.4 that are not in the tables. The supplement also contains an independent re-implementation in base R (R 4.3), written from this paper rather than translated from Python; it simulates its own counts and is provided as a check on the reported effects, not as a second route to the published numbers. The supplement is archived at Zenodo, https://doi.org/10.5281/zenodo.23459961; any later revision of it will be deposited there as a new version.

Conflict of interest

The authors declare no conflict of interest.

Appendix

A. Exact mean and variance of the two-length mixture model

For completeness, we record the exact finite-n moments of the two-length mixture model [6, 7]. By Lemma 2, the formulas are valid under either seeding convention, provided $\mathbb{E}(L_g)$ and $\mathrm{Var}(L_g)$ are evaluated under the seeding region actually used for type g. With $p^{(gg')} = 2\mathbb{E}(L_g)\mathbb{E}(L_{g'})/(\pi|C|)$,

$$\mathbb{E}(Y) = \binom{n_1}{2} p^{(11)} + \binom{n_2}{2} p^{(22)} + n_1 n_2\, p^{(12)},$$

and, writing $\varsigma_{gg'} := 4\mathbb{E}(L_g)\mathbb{E}(L_{g'})/(\pi^2|C|^2)$,

$$\begin{aligned}\mathrm{Var}(Y) = {} & \binom{n_1}{2} p^{(11)}(1-p^{(11)}) + \binom{n_2}{2} p^{(22)}(1-p^{(22)}) + n_1 n_2\, p^{(12)}(1-p^{(12)}) \\ & + \varsigma_{11}\,\mathrm{Var}(L_1)\, n_1(n_1-1)(n_1-2) + \varsigma_{22}\,\mathrm{Var}(L_2)\, n_2(n_2-1)(n_2-2) \\ & + \varsigma_{22}\,\mathrm{Var}(L_1)\, n_1 n_2(n_2-1) + \varsigma_{11}\,\mathrm{Var}(L_2)\, n_1(n_1-1) n_2 \\ & + 2\varsigma_{12}\,\mathrm{Var}(L_1)\, n_1(n_1-1) n_2 + 2\varsigma_{12}\,\mathrm{Var}(L_2)\, n_1 n_2(n_2-1),\end{aligned}$$

where the three terms on the first line are the Bernoulli diagonal, of order n^2, and the six that follow are the covariance contributions of order n^3, corresponding to the enumeration in [7, Table 5].

The same bookkeeping extends to G fiber types. The mean has one term per unordered pair of types, $O(G^2)$ in number, and the variance adds one covariance term per ordered triple (type of the shared fiber, types of the two partners), $O(G^3)$ in number. Every term is a product of a combinatorial count with closed-form constants built from $\mathbb{E}(L_g), \mathrm{Var}(L_g)$, so the evaluation of (m, v) costs a few hundred arithmetic operations even for a dozen types and is instantaneous. The cost of the exercise lies entirely in the truncated-length moments, one pair $(\mathbb{E}(L_g), \mathrm{Var}(L_g))$ per type, and those are closed forms by Lemma 1.

References

1. O. Kallmes, H. Corte, The structure of paper. I. The statistical geometry of an ideal two dimensional fiber network, *Tappi J.*, **43** (1960), 737–752.
2. T. Komori, Y. Ujihara, Y. Matsunaga, K. Makishima, Crossings of curled fibers in two-dimensional assemblies, *Tappi*, **62** (1979), 93–95.
3. Y. B. Yi, L. Berhan, A. M. Sastry, Statistical geometry of random fibrous networks, revisited: Waviness, dimensionality, and percolation, *J. Appl. Phys.*, **96** (2004), 1318–1327. https://doi.org/10.1063/1.1763240
4. M. W. Suh, H. Chun, R. L. Berger, P. Bloomfield, Distribution of fiber intersections in two-dimensional random fiber webs–A basic geometrical probability model, *Text. Res. J.*, **80** (2010), 301–311. https://doi.org/10.1177/0040517509105071
5. H. Chun, Probabilistic modeling of fiber length segments within a bounded area of two-dimensional fiber webs, *Commun. Stat. Appl. Met.*, **18** (2011), 301–317. https://doi.org/10.5351/CKSS.2011.18.3.301

6. H. Chun, M. W. Suh, Distribution of fiber intersections in two-dimensional random fiber web cases with a mixture of two fiber lengths, *Text. Res. J.*, **90** (2020), 1851–1859. https://doi.org/10.1177/0040517519898158
7. H. Chun, J. H. Jhong, Asymptotic distribution of fiber intersection counts in two-dimensional random fiber webs, *AIMS Math.*, **11** (2026), 15008–15027. https://doi.org/10.3934/math.2026617
8. Y. Wang, B. Zhou, C. Fu, X. Xu, Modeling cure fractions in interval-censored competing risks data: An AFT mixture approach, *Electron. Res. Arch.*, **34** (2026), 6573–6613. https://doi.org/10.3934/era.2026288
9. P. C. Consul, G. C. Jain, A generalization of the Poisson distribution, *Technometrics*, **15** (1973), 791–799. https://doi.org/10.1080/00401706.1973.10489112
10. R. W. Conway, W. L. Maxwell, A queuing model with state dependent service rates, *J. Ind. Eng.*, **12** (1962), 132–136.
11. G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, P. Boatwright, A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution, *J. R. Stat. Soc. C-Appl.*, **54** (2005), 127–142. https://doi.org/10.1111/j.1467-9876.2005.00474.x
12. A. D. Barbour, L. Holst, S. Janson, *Poisson approximation*, Oxford: Oxford University Press, 1992. https://doi.org/10.1093/oso/9780198522355.001.0001
13. M. H. Quenouille, A relation between the logarithmic, Poisson, and negative binomial series, *Biometrics*, **5** (1949), 162–164. https://doi.org/10.2307/3001917
14. N. L. Johnson, A. W. Kemp, S. Kotz, *Univariate discrete distributions*, 3 Eds., Hoboken: Wiley, 2005. https://doi.org/10.1002/0471715816
15. F. Ouimet, A refined continuity correction for the negative binomial distribution and asymptotics of the median, *Metrika*, **86** (2023), 827–849. https://doi.org/10.1007/s00184-023-00897-2
16. T. C. Brown, M. J. Phillips, Negative binomial approximation with Stein's method, *Methodol. Comput. Appl.*, **1** (1999), 407–421. https://doi.org/10.1023/A:1010094221471
17. A. D. Barbour, H. L. Gan, A. Xia, Stein factors for negative binomial approximation in Wasserstein distance, *Bernoulli*, **21** (2015), 1002–1013. https://doi.org/10.3150/14-BEJ595
18. Y. D. Liang, B. A. Barsky, A new concept and method for line clipping, *ACM T. Graphic.*, **3** (1984), 1–22. https://doi.org/10.1145/357332.357333
19. D. S. Dimitrova, V. K. Kaishev, S. Tan, Computing the Kolmogorov–Smirnov distribution when the underlying cdf is purely discrete, mixed, or continuous, *J. Stat. Softw.*, **95** (2020), 1–42. https://doi.org/10.18637/jss.v095.i10
20. H. Callaert, P. Janssen, The Berry–Esseen theorem for U-statistics, *Ann. Stat.*, **6** (1978), 417–421. https://doi.org/10.1214/aos/1176344132
21. V. S. Korolyuk, Y. V. Borovskich, *Theory of U-statistics*, In: Mathematics and Its Applications, Dordrecht: Springer, 1994. https://doi.org/10.1007/978-94-017-3515-5
22. P. J. Bickel, F. Götze, W. R. V. Zwet, The Edgeworth expansion for U-statistics of degree two, *Ann. Stat.*, **14** (1986), 1463–1484. https://doi.org/10.1214/aos/1176350170

23. V. Bentkus, F. Götze, W. R. V. Zwet, An Edgeworth expansion for symmetric statistics, *Ann. Stat.*, **25** (1997), 851–896. https://doi.org/10.1214/aos/1031833676

24. W. Wang, H. Wu, Z. Dai, W. Liu, Distributed optimal subsampling inference for the additive hazards model with massive survival data, *BMC Med. Res. Methodol.*, 2026. https://doi.org/10.1186/s12874-026-02985-z



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)