

Session 6: Continuous Random Variables

Movement II: One Random Variable

Instructor: John Lipor

1 Launch

Your range sensor returns 103.7 m. Ask for the probability that the true range is *exactly* 103.7 m and the answer is zero: no single real number carries any probability when the outcome varies continuously. Yet the reading is obviously informative, and you would happily bet on the target being nearer to 103.7 m than to 200 m. Resolving that tension is today's job. The fix is to stop asking for the probability of a point and start asking for the probability of an *interval*, and the object that delivers interval probabilities is the density, a rate of probability per unit length rather than a probability itself. By the end you will read a Gaussian noise spec, standardize it, and compute an outage probability from a table; and you will see why two noise models with the *same variance* can still disagree, sharply, about how alarming a large reading is.

2 Development

2.1 From CDF to density

Session 4 built the CDF $F_X(x) = \mathbb{P}(X \leq x)$ for discrete variables, where it was a staircase. The CDF is the object that survives into the continuous world unchanged in meaning: for *any* random variable, $F_X(x)$ is the probability of landing at or below x . What changes is that a continuous variable's CDF has no jumps; it rises smoothly, and the mass it accumulates over an interval is what we can measure.

Definition 1 (Continuous random variable and its density). *A random variable X is **continuous** if its CDF can be written as an integral of a nonnegative function f_X , called the **probability density function (PDF)**:*

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt. \quad (1)$$

Consequently $f_X(x) = F'_X(x)$ wherever F_X is differentiable, and the probability of an interval is the area under the density:

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(t) dt = F_X(b) - F_X(a). \quad (2)$$

Because probability comes only from intervals, a single point carries none: taking $a \rightarrow b$ in (2) gives $\mathbb{P}(X = b) = 0$ for every b . This is the resolution of the Launch puzzle. It also means that for continuous variables we need not fuss over $<$ versus $\leq$; endpoints have zero probability either way.

A density is a rate, not a probability. The value $f_X(x)$ is *not* a probability. It is a probability *per unit length*: the mass in a sliver of width dx at x is $f_X(x) dx$. Two consequences that trip people up. First, a density may exceed 1 (a Uniform $[0, \frac{1}{2}]$ has $f = 2$ on its support) because it is a rate, and only its integral must be 1. Second, $f_X(x) = 0.02$ does not mean “a 2% chance”; it means nothing at all until multiplied by a length. This resolves the Session 1 worry about conditioning on a probability-zero event, which we will complete in Session 8: we condition on $\{X = x\}$ through the density, not through a probability.

2.2 Expectation and the tail-integral formula

The mean and LOTUS carry over from the discrete case with sums replaced by integrals.

Definition 2 (Expectation). *For a continuous random variable X with density f_X ,*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx, \quad (3)$$

and more generally, by LOTUS, $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$ for any function g (we assert the continuous LOTUS; the proof mirrors the discrete regrouping of Session 4).

Session 5 met a discrete tail-sum formula and deferred its proof. Its continuous twin we prove now, because the proof is short and the identity is the workhorse for means of nonnegative quantities.

Theorem 1 (Tail-integral formula). *For a nonnegative random variable X ,*

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x) dx. \quad (4)$$

Proof. For any fixed value $X \geq 0$, height and area agree: $X = \int_0^{\infty} \mathbb{1}\{X > x\} dx$, since the integrand is 1 for $x < X$ and 0 for $x > X$. Take the expectation of both sides and exchange it with the integral over x (a step we assert here; it is justified because the integrand is nonnegative):

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^{\infty} \mathbb{1}\{X > x\} dx\right] = \int_0^{\infty} \mathbb{E}[\mathbb{1}\{X > x\}] dx = \int_0^{\infty} \mathbb{P}(X > x) dx, \quad (5)$$

using $\mathbb{E}[\mathbb{1}\{A\}] = \mathbb{P}(A)$ from Session 4 at the last step. □

This is Session 5’s tail sum wearing continuous clothes: there we added $\mathbb{P}(X \geq k)$ over integers, here we integrate $\mathbb{P}(X > x)$ over the half-line. Same idea, same picture, integral instead of sum.

2.3 Three models: uniform, exponential, Gaussian

Uniform. $X \sim \text{Uniform}[a, b]$ spreads its mass evenly:

$$f_X(x) = \frac{1}{b-a} \text{ on } [a, b], \quad F_X(x) = \frac{x-a}{b-a} \text{ on } [a, b]. \quad (6)$$

By symmetry $\mathbb{E}[X] = (a+b)/2$, and a short computation gives $\text{Var}(X) = (b-a)^2/12$. It is the model for “somewhere in this interval, no place preferred.”

Exponential. $X \sim \text{Exponential}(\lambda)$, parameterized by *rate* $\lambda > 0$, is the standard model for a nonnegative waiting time:

$$f_X(x) = \lambda e^{-\lambda x}, \quad F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0. \quad (7)$$

Its survival probability is $\mathbb{P}(X > x) = e^{-\lambda x}$, and the tail-integral formula delivers the mean in one line:

$$\mathbb{E}[X] = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}. \quad (8)$$

The exponential is the continuous echo of Session 5’s geometric, and it inherits the geometric’s signature property.

Proposition 1 (Memorylessness of the exponential). *If $X \sim \text{Exponential}(\lambda)$, then for all $s, t \geq 0$,*

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t). \quad (9)$$

Proof. Since $\{X > s+t\} \subseteq \{X > s\}$, the intersection is $\{X > s+t\}$, so

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t). \quad (10)$$

□

A component that has already survived s hours is, going forward, as good as new: its remaining lifetime has the same distribution it started with. This is the exact continuous mirror of the geometric argument from Session 5, and it is no accident. A geometric with small per-step success probability q , counted on a fine time grid of spacing Δ with $\lambda = q/\Delta$, converges to an $\text{Exponential}(\lambda)$ as $\Delta \rightarrow 0$; the memoryless discrete clock becomes the memoryless continuous one. And the same reconciliation from Session 5 applies: memorylessness is a statement about the waiting time given the mechanism, not a license to ignore evidence about whether the mechanism is even running.

Gaussian. $X \sim \mathcal{N}(\mu, \sigma^2)$, the workhorse noise model, has density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad (11)$$

with mean μ and variance σ^2 . We take the normalizing constant $1/\sqrt{2\pi\sigma^2}$ on faith (verifying $\int f_X = 1$ is a standard polar-coordinates exercise). The Gaussian has no elementary antiderivative, so we never integrate it by hand; instead we *standardize* to a single reference variable and read a table.

Definition 3 (Standard normal, Φ , and Q). *The standard normal is $Z \sim \mathcal{N}(0, 1)$. Its CDF is written $\Phi(z) = \mathbb{P}(Z \leq z)$, and its upper-tail probability is*

$$Q(z) = \mathbb{P}(Z > z) = 1 - \Phi(z). \quad (12)$$

By the symmetry of the density about 0, $Q(z) = \Phi(-z)$ and $\Phi(z) = 1 - Q(z)$.

The move that makes every Gaussian computation identical is standardization: if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$, so any probability about X becomes a probability about Z :

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (13)$$

Note the divisor is σ , the standard deviation, never σ^2 . A few tail values, worth memorizing to one or two figures, anchor all Gaussian estimates:

z	$Q(z) = \mathbb{P}(Z > z)$
1	0.1587
2	0.0228
3	0.00135

The “ 3σ ” folklore is the last row: only about one part in 740 of a Gaussian’s mass lies beyond three standard deviations on one side.

Why Gaussian noise models are everywhere. The central limit theorem says that if $X_1, X_2, \dots$ are iid with mean μ and variance σ^2 , then the standardized sum $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu)$ converges in distribution to $\mathcal{N}(0, 1)$ as $n \rightarrow \infty$. Because physical noise is so often the accumulation of many small independent disturbances, the Gaussian is its natural limiting shape, which is why it is the default sensor-noise model throughout this course (we state the theorem here and return to sums of Gaussians in Session 14).

2.4 Tails are a modeling decision

Variance is not the whole story of a noise model. Two densities can share a mean and a variance and still disagree about how often they produce a large deviation, and that disagreement is exactly what decides whether a surprising reading is “just noise” or a broken sensor. To see it, put the Gaussian beside the **Laplace** density (a double-sided exponential), used here *only* as a comparison model:

$$f(z) = \frac{\lambda}{2} e^{-\lambda|z|}, \quad \text{Var} = \frac{2}{\lambda^2}. \quad (14)$$

Match its variance to a $\mathcal{N}(0, \sigma^2)$ by setting $2/\lambda^2 = \sigma^2$, i.e. $\lambda = \sqrt{2}/\sigma$. Now both models have the same spread, and we can ask each for the probability of landing more than k standard deviations from the mean. For the Gaussian, $\mathbb{P}(|X| > k\sigma) = 2Q(k)$; for the variance-matched Laplace, $\mathbb{P}(|X| > k\sigma) = e^{-\sqrt{2}k}$:

k	Gaussian $2Q(k)$	Laplace $e^{-\sqrt{2}k}$
1	0.317	0.243
2	0.045	0.059

Near the center ($k = 1$) the two are comparable and the Gaussian actually puts slightly more mass beyond 1σ . But the ordering has already reversed by $k = 2$, and the reason is structural: the Gaussian tail decays like $e^{-k^2/2}$, quadratically fast in the exponent, while the Laplace decays only like $e^{-\sqrt{2}k}$, linearly. The heavier (Laplace) tail pulls steadily ahead, and the gap *widens* as k grows. How much worse it gets at 3σ and 4σ , and what that means for judging whether a 4σ reading is noise or a broken sensor, is the subject of HW3.

Same variance $\neq$ same model. Matching variances does not make two noise models interchangeable. The tail behavior, how fast the density decays far from the mean, is a separate modeling choice, and it is the one that governs rare-but-consequential events: outages, false alarms, and the question “is this outlier real?” Choose the tail deliberately.

3 Worked examples

3.1 (a) Outage probability from a Gaussian spec

A receiver’s received power, in decibels, is modeled as $X \sim \mathcal{N}(-70, 36)$, so the mean is -70 dB and the standard deviation is $\sigma = \sqrt{36} = 6$ dB. The receiver loses lock (an *outage*) whenever the power drops below -82 dB. What is the outage probability?

Standardize with $\sigma = 6$ (not 36), using (13):

$$\mathbb{P}(\text{outage}) = \mathbb{P}(X < -82) = \Phi\left(\frac{-82 - (-70)}{6}\right) = \Phi\left(\frac{-12}{6}\right) = \Phi(-2). \quad (15)$$

By symmetry $\Phi(-2) = Q(2)$, and the table gives

$$\mathbb{P}(\text{outage}) = Q(2) \approx 0.0228, \quad (16)$$

about a 2.3% chance. The entire computation was: subtract the mean, divide by the standard deviation, read the table. That is every Gaussian probability you will ever compute.

3.2 (b) First alarm among four sensors

$K = 4$ sensors watch a field; once a target enters, each raises an alarm after an independent $\text{Exponential}(\lambda)$ delay. When does the *first* alarm sound?

The first alarm time is $T = \min(X_1, \dots, X_4)$, and the minimum is the moment for Session 5’s CDF method. The first alarm sounds after t exactly when *all four* sensors are still silent:

$$\mathbb{P}(T > t) = \mathbb{P}(X_1 > t, \dots, X_4 > t) = \prod_{i=1}^4 \mathbb{P}(X_i > t) = \left(e^{-\lambda t}\right)^4 = e^{-4\lambda t}. \quad (17)$$

But $e^{-4\lambda t}$ is the survival probability of an $\text{Exponential}(4\lambda)$, so $T \sim \text{Exponential}(4\lambda)$: the minimum of independent exponentials is again exponential, with the *rates added*. Four sensors detect four times as fast, with mean first-alarm time $1/(4\lambda)$. Notice this reused Session 5’s recipe unchanged, only swapping the geometric’s $(1 - q)^k$ for the exponential’s $e^{-\lambda t}$; the min-means-CDF lesson does not care whether the clock is discrete or continuous.

3.3 (c) The last alarm, and a new named distribution

Return to the sensor field of example (b), but now ask about the *last* alarm. A confirmation protocol declares a target only once every one of m sensors has reported, and each sensor's delay is an independent $\text{Exponential}(\lambda)$. Measure time in units of the mean delay, so $\lambda = 1$ (for general λ , replace M by λM throughout). How long does confirmation take?

The confirmation time is $M = \max(X_1, \dots, X_m)$, so Session 5's recipe applies with the maximum in place of the minimum:

$$\mathbb{P}(M \leq t) = \prod_{i=1}^m \mathbb{P}(X_i \leq t) = (1 - e^{-t})^m, \quad t \geq 0. \quad (18)$$

That answers the question exactly, but it hides how M behaves as the field grows. The first alarm got faster, like $1/m$. The last alarm gets slower, and a good guess for the rate is $\ln m$: the waiting time t at which a single sensor is still silent with probability $1/m$ solves $e^{-t} = 1/m$. So center M at $\ln m$ and look at what is left over. For $x \geq -\ln m$,

$$\mathbb{P}(M - \ln m \leq x) = (1 - e^{-(x + \ln m)})^m = \left(1 - \frac{e^{-x}}{m}\right)^m. \quad (19)$$

This is the same limit that produced the Poisson pmf in Session 5, $(1 + a/m)^m \rightarrow e^a$, now with $a = -e^{-x}$ held fixed while $m \rightarrow \infty$:

$$\mathbb{P}(M - \ln m \leq x) \longrightarrow \exp(-e^{-x}) \quad \text{for every real } x. \quad (20)$$

The shifted maximum does not spread out or drift away. It settles on a fixed shape, and that shape deserves a name.

Definition 4 (Gumbel distribution). *A random variable G has the **standard Gumbel distribution** if its CDF is*

$$F(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}. \quad (21)$$

*It has the **Gumbel distribution with location** μ , written $G \sim \text{Gumbel}(\mu)$, if $G - \mu$ is standard Gumbel, so that $\mathbb{P}(G \leq x) = \exp(-e^{-(x-\mu)})$.*

Check that F is a legitimate CDF: it is continuous and increasing (because e^{-x} decreases and $\exp(-\cdot)$ reverses that), $F(x) \rightarrow \exp(-\infty) = 0$ as $x \rightarrow -\infty$, and $F(x) \rightarrow \exp(0) = 1$ as $x \rightarrow \infty$. Differentiating gives the density $f(x) = e^{-x} \exp(-e^{-x})$. Unlike the exponential, the Gumbel lives on the whole real line; it is skewed, with a left tail that dies doubly exponentially and a right tail that dies like e^{-x} . In this language, (20) says that $M - \ln m$ is approximately standard Gumbel; equivalently, the last alarm M itself is approximately $\text{Gumbel}(\ln m)$.

How fast is the limit reached? At $x = 0$ the exact probability $(1 - 1/m)^m$ is 0.316 for $m = 4$, 0.358 for $m = 20$, and 0.366 for $m = 100$, against the limit $e^{-1} \approx 0.368$ (Figure 1). Twenty sensors are already plenty.

Designing the timeout. With $m = 20$ sensors, how long must the protocol wait so that all twenty have reported with probability 0.95? The Gumbel approximation asks for $\exp(-e^{-x}) = 0.95$, so $x = -\ln(-\ln 0.95) \approx 2.970$ and $t \approx \ln 20 + 2.970 \approx 5.966$ mean

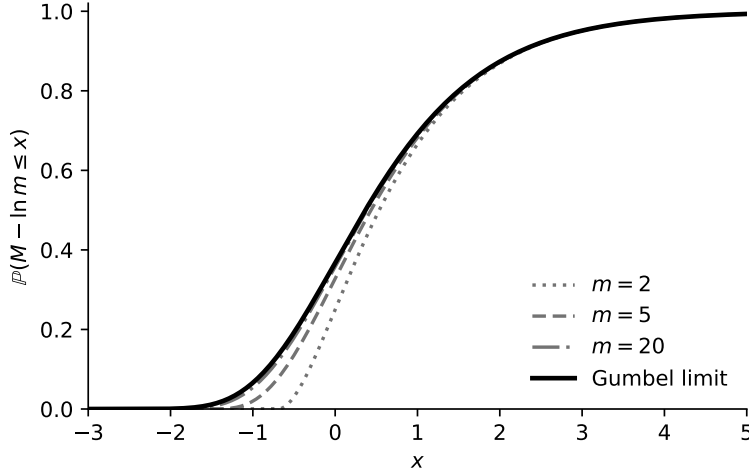


Figure 1: The exact CDF of $M - \ln m$ from (18) for $m = 2, 5, 20$ approaches the Gumbel CDF (20).

delays. The exact answer, solving $(1 - e^{-t})^{20} = 0.95$, is $t = -\ln(1 - 0.95^{1/20}) \approx 5.967$. The approximation also says how the timeout scales: doubling the field to 40 sensors adds only $\ln 2 \approx 0.69$ mean delays. Detection got four times faster with four sensors in example (b); confirmation gets only logarithmically slower as sensors are added.

Remark 1. *The Gumbel shape is not special to exponentials. Maxima of many light-tailed families, the Gaussian included, approach it after a suitable shift and scaling, which is why it is called an extreme-value distribution. We will not need that general fact.*

3.4 (d) Maxima of Gumbels are Gumbel

The minimum of independent exponentials was again exponential, with rates added. The Gumbel has the mirror-image property for maxima.

Proposition 2 (Max-stability). *Let $G_1, \dots, G_n$ be independent with $G_k \sim \text{Gumbel}(\mu_k)$. Then*

$$\max_k G_k \sim \text{Gumbel}(L), \quad L = \ln \sum_{k=1}^n e^{\mu_k}. \quad (22)$$

In particular, the maximum of n independent $\text{Gumbel}(\mu)$ variables is $\text{Gumbel}(\mu + \ln n)$.

Proof. Session 5's proof never used that the CDFs were identical: the maximum is at most x exactly when every G_k is, and independence turns that into a product. So

$$\mathbb{P}\left(\max_k G_k \leq x\right) = \prod_{k=1}^n \exp\left(-e^{-(x-\mu_k)}\right) = \exp\left(-e^{-x} \sum_{k=1}^n e^{\mu_k}\right) = \exp\left(-e^{-(x-L)}\right), \quad (23)$$

using $\sum_k e^{\mu_k} = e^L$ in the last step. This is the $\text{Gumbel}(L)$ CDF. With all $\mu_k = \mu$, $L = \ln(ne^\mu) = \mu + \ln n$. $\square$

The parallel with the exponential is exact: for minima of exponentials the rates λ_k add; for maxima of Gumbels the weights e^{μ_k} add. Session 7 will show this is not a coincidence.

The Proposition also explains why example (c)’s limit had to be Gumbel. Split mn sensors into n blocks of m . Each block’s last alarm is approximately Gumbel($\ln m$), so by the Proposition the overall last alarm is approximately Gumbel($\ln m + \ln n$) = Gumbel($\ln mn$), which is exactly what (20) predicts for mn sensors directly. A limit for maxima has to survive being maximized again, and the Gumbel does.

Log-sum-exp is a soft maximum. The location $L = \ln \sum_k e^{\mu_k}$ of the maximum satisfies

$$\max_k \mu_k \leq L \leq \max_k \mu_k + \ln n, \quad (24)$$

because the sum $\sum_k e^{\mu_k}$ lies between its largest term and n times its largest term. So L is a smoothed version of $\max_k \mu_k$: the largest location dominates, and close competitors push the maximum a little higher. Session 9 will turn this soft maximum into the softmax.

4 Consolidation

Work these in class.

1. (*Gaussian standardization.*) A range estimate is $X \sim \mathcal{N}(50, 4)$ (so $\sigma = 2$ m). Find $\mathbb{P}(X > 53)$.
Answer: standardize with $\sigma = 2$: $\mathbb{P}(X > 53) = Q((53 - 50)/2) = Q(1.5) \approx 0.0668$.
2. (*Exponential, memorylessness.*) An alarm delay is Exponential(λ) with mean 5 s, so $\lambda = 0.2$. Compute $\mathbb{P}(\text{delay} > 10)$, and then $\mathbb{P}(\text{delay} > 15 \mid \text{delay} > 5)$.
Answer: $\mathbb{P}(\text{delay} > 10) = e^{-0.2 \cdot 10} = e^{-2} \approx 0.135$; by memorylessness $\mathbb{P}(\text{delay} > 15 \mid \text{delay} > 5) = \mathbb{P}(\text{delay} > 10) \approx 0.135$.
3. (*Conceptual, density vs probability.*) A colleague objects that a Uniform $[0, \frac{1}{2}]$ density equals 2, “which is more than 100%, so it cannot be right.” In one or two sentences, resolve the objection.
Answer: $f_X(x) = 2$ is a rate (probability per unit length), not a probability; the total probability is the area $2 \times \frac{1}{2} = 1$, and densities are unbounded in principle because only their integral must equal one.

5 Synthesis

The course thesis asks “what is uncertain?” and this session answered it for quantities that vary continuously: the CDF, unchanged in meaning from the discrete case, is the honest object, and the density beneath it is a *rate* of probability per unit length, which is why a single point carries no probability yet an interval does. From that one shift everything followed. The tail-integral formula gave means of nonnegative quantities directly from survival probabilities, the continuous echo of Session 5’s tail sum. The exponential turned out to be the geometric’s memoryless twin; the *last* of many exponential alarms, recentered at $\ln m$, produced a new

named law, the Gumbel, which is closed under maxima just as the exponential is closed under minima; and the Gaussian reduced every question to “subtract the mean, divide by the standard deviation, read Q .” The session’s sharpest lesson was that variance does not determine a model: the Gaussian and the variance-matched Laplace agree on spread yet disagree on tails, and the tail is the modeling decision that governs rare, expensive events. Next session pushes a random variable *through a function*, deriving the density of $Y = g(X)$, which is how instruments and simulators are actually modeled, and which finally explains what happens each time you call a library sampler.

Self-check questions

1. A continuous X has $f_X(103.7) = 0.02$. Is the probability that X equals 103.7 therefore 0.02? If not, what does the number mean?

Answer: No; $\mathbb{P}(X = 103.7) = 0$ for any continuous X . The value 0.02 is a density (probability per unit length): the mass near 103.7 in a window of width dx is about $0.02 dx$.

2. For $X \sim \mathcal{N}(\mu, \sigma^2)$, write $\mathbb{P}(X > c)$ in terms of Q . What is the single most common arithmetic slip in this step?

Answer: $\mathbb{P}(X > c) = Q((c - \mu)/\sigma)$; the common slip is dividing by the variance σ^2 instead of the standard deviation σ .

3. Two zero-mean noise models have the same variance, one Gaussian and one Laplace. In one sentence, can you conclude they assign the same probability to a large outlier, and why or why not?

Answer: No; equal variance fixes the spread but not the tail, and the heavier-tailed Laplace assigns more probability to large deviations, with the gap growing as the deviation grows.

4. With $m = 10$ sensors and $\lambda = 1$, use the Gumbel approximation to estimate the median confirmation time. (The standard Gumbel median is $-\ln \ln 2 \approx 0.367$.) Then find the exact median from (18).

Answer: approximately $\ln 10 + 0.367 \approx 2.669$; exactly $-\ln(1 - 0.5^{1/10}) \approx 2.704$.

5. Why is the location of the maximum of n independent Gumbel(μ) variables $\mu + \ln n$, rather than μ or $n\mu$?

Answer: The maximum is at least as large as any one variable, so its location cannot fall below μ ; but log-sum-exp exceeds the largest location by at most $\ln n$, and with n equal locations it attains that bound exactly, since $\ln(ne^\mu) = \mu + \ln n$. More competitors push the maximum up, but only logarithmically.