Continuous Distributions

1.8-1.9: Continuous Random Variables

1.10.1: Uniform Distribution (Continuous)

1.10.4-5 Exponential and Gamma Distributions: Distance between crossovers

Prof. Tesler

Math 283
Fall 2019

Cumulative Distribution Function (CDF)

Cumulative Distribution Function (CDF)

Discrete random variables

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

- The *Cumulative Distribution Function (CDF)* of random variable X is
$$F_X(x) = P(X \leqslant x)$$

- $$F_X(1.5) = P(X \leqslant 1.5) = P_X(0.5) + P_X(1.0) + P_X(1.5) = 0.1 + 0.2 + 0.3 = 0.6$$

- **In-between points with nonzero probability:**
$$F_X(1.7) = P(X \leqslant 1.7) = P(X \leqslant 1.5) = F_X(1.5) = 0.6$$
whereas the PDF there is 0: $\quad P_X(1.7) = 0$

- Similarly, $F_X(k) = F_X(1.5) = 0.6$ for $1.5 \leqslant k < 2.0$.

CDF outside of the range

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

- $F_X(-1) = P(X \leqslant -1) = 0$ (no points w/nonzero PDF)
- $F_X(5) = P(X \leqslant 5) = 1$ (has all of the points w/nonzero PDF)

General case

$$\lim_{k \to -\infty} F_X(k) = 0 \qquad \lim_{k \to +\infty} F_X(k) = 1$$

CDF table

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

CDF

k	$F_X(k)$
$k < 0.5$	0
$0.5 \leqslant k < 1.0$	0.1
$1.0 \leqslant k < 1.5$	0.3
$1.5 \leqslant k < 2.0$	0.6
$2.0 \leqslant k < 2.5$	0.7
$2.5 \leqslant k < 3.0$	0.8
$3.0 \leqslant k$	1

Using CDF table with various inequalities: $\leqslant$, $>$, $<$, $\geqslant$

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

CDF

k	$F_X(k)$
$k < 0.5$	0
$0.5 \leqslant k < 1.0$	0.1
$1.0 \leqslant k < 1.5$	0.3
$1.5 \leqslant k < 2.0$	0.6
$2.0 \leqslant k < 2.5$	0.7
$2.5 \leqslant k < 3.0$	0.8
$3.0 \leqslant k$	1

- $P(X \leqslant 1) = 0.3$
- $P(X > 1) = 1 - P(X \leqslant 1) = 0.7$
- $P(X < 1) = P(X \leqslant 1^-) = F_X(1^-) = 0.1$
  using infinitesimal notation from Calculus: 1^- is just below 1, like 0.99999999, but even closer.
- $P(X \geqslant 1) = 1 - P(X < 1) = 1 - F_X(1^-) = 0.9$

Using CDF table on an interval

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

CDF

k	$F_X(k)$
$k < 0.5$	0
$0.5 \leqslant k < 1.0$	0.1
$1.0 \leqslant k < 1.5$	0.3
$1.5 \leqslant k < 2.0$	0.6
$2.0 \leqslant k < 2.5$	0.7
$2.5 \leqslant k < 3.0$	0.8
$3.0 \leqslant k$	1

$$F_X(2) = P(X \leqslant 2) = P_X(0.5) + P_X(1.0) + P_X(1.5) + P_X(2.0)$$
$$F_X(1) = P(X \leqslant 1) = P_X(0.5) + P_X(1.0)$$

$$\begin{aligned} P(1 < X \leqslant 2) &= P_X(1.5) + P_X(2.0) \\ &= P(X \leqslant 2) - P(X \leqslant 1) = F_X(2) - F_X(1) \\ &= 0.7 - 0.3 = 0.4 \end{aligned}$$

Converting intervals to the form $P(a < X \leqslant b)$

PDF

k	$P_X(k)$
0.5	0.1
1.0	0.2
1.5	0.3
2.0	0.1
2.5	0.1
3.0	0.2

CDF

k	$F_X(k)$
$k < 0.5$	0
$0.5 \leqslant k < 1.0$	0.1
$1.0 \leqslant k < 1.5$	0.3
$1.5 \leqslant k < 2.0$	0.6
$2.0 \leqslant k < 2.5$	0.7
$2.5 \leqslant k < 3.0$	0.8
$3.0 \leqslant k$	1

The formula $P(a < X \leqslant b) = F_X(b) - F_X(a)$ uses $a < X$ (not $a \leqslant X$) and $X \leqslant b$ (not $X < b$). Other formats must be converted to this:

- $P(1 < X \leqslant 2) = F_X(2) - F_X(1) = 0.7 - 0.3 = 0.4$
- $P(1 \leqslant X \leqslant 2) = P(1^- < X \leqslant 2) = F_X(2) - F_X(1^-) = 0.7 - 0.1 = 0.6$
- $P(1 < X < 2) = P(1 < X \leqslant 2^-) = F_X(2^-) - F_X(1) = 0.6 - 0.3 = 0.3$
- $P(1 \leqslant X < 2) = P(1^- < X \leqslant 2^-) = F_X(2^-) - F_X(1^-) = 0.6 - 0.1 = 0.5$

Continuous distributions

Continuous distributions

Example

- Pick a real number x between 20 and 30 with all real values in $[20, 30]$ equally likely.
- Sample space: $S = [20, 30]$
- Number of outcomes: $|S| = \infty$
- Probability of each outcome: $P(X = x) = \frac{1}{\infty} = 0$
- Yet, $P(X \leqslant 21.5) = 15\%$

Continuous distributions

- The *sample space* S is often a subset of $\mathbb{R}^n$.
  We'll do the 1-dimensional case $S \subset \mathbb{R}$.
- The *probability density function (PDF)* $f_X(x)$ is defined differently than the discrete case:
  - $f_X(x)$ is a real-valued function on S with $f_X(x) \geqslant 0$ for all $x \in S$.
  - $\int_S f_X(x)\,dx = 1$ (vs. $\sum_{x \in S} P_X(x) = 1$ for discrete)
  - The probability of event $A \subset S$ is $P(A) = \int_A f_X(x)\,dx$ (vs. $\sum_{x \in A} P_X(x)$).
  - In n dimensions, use n-dimensional integrals instead.
  - **Notation:** Uppercase F for CDF vs. lowercase f for pdf.

Uniform distribution

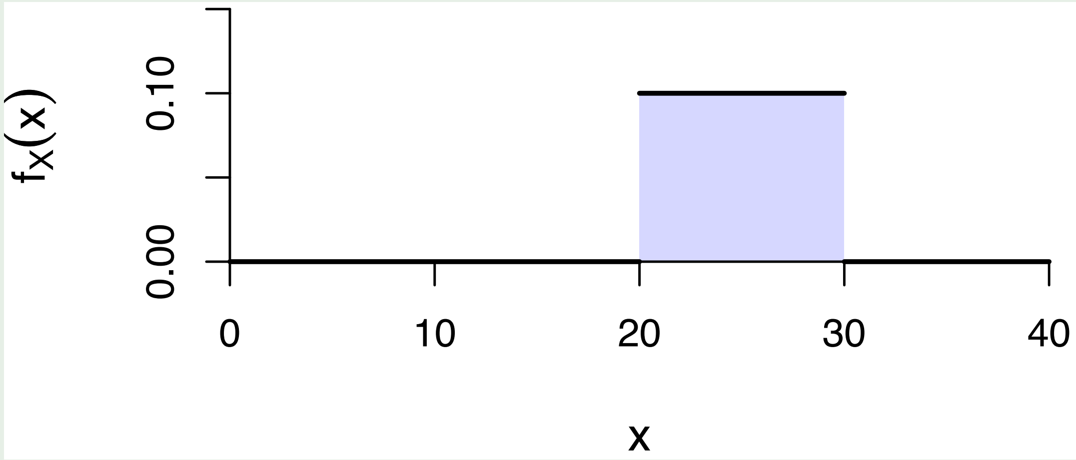
- Let $a < b$ be real numbers.
- The *Uniform Distribution* on $[a, b]$ is that all numbers in $[a, b]$ are "equally likely."
- More precisely, $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leqslant x \leqslant b; \\ 0 & \text{otherwise.} \end{cases}$

Uniform distribution (real case)

The uniform distribution on $[20, 30]$

We could regard the sample space as $[20, 30]$, or as all reals.

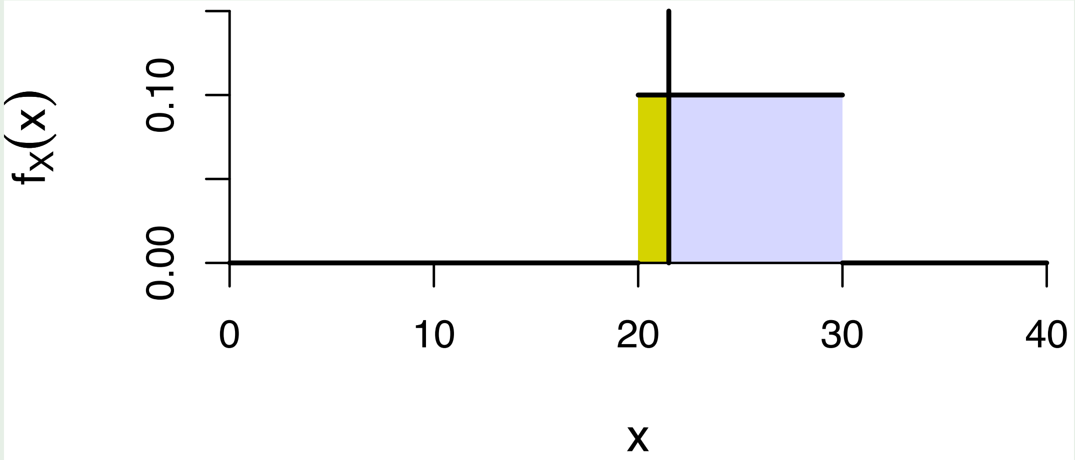
$$f_X(x) = \begin{cases} 1/10 & \text{for } 20 \leqslant x \leqslant 30; \\ 0 & \text{otherwise.} \end{cases}$$



$$P(X \leqslant 21.5) = \int_{-\infty}^{20} 0\, dx + \int_{20}^{21.5} \frac{1}{10} dx = 0 + \left. \frac{x}{10} \right|_{20}^{21.5}$$

$$= \frac{21.5 - 20}{10}$$

$$= .15 = 15\%$$



Cumulative distribution function (CDF)

The *Cumulative Distribution Function (CDF)* of a random variable X is

$$F_X(x) = P(X \leqslant x)$$

- For a continuous random variable,

  $$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} f_X(t)\,dt \qquad \text{and} \qquad f_X(x) = F_X{}'(x)$$

- The integral cannot have "x" as the name of the variable in both of $F_X(x)$ and $f_X(x)$ because one is the upper limit of the integral and the other is the integration variable. So we use two variables x, t.
- We can either write

  $$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} f_X(t)\,dt$$

  or

  $$F_X(t) = P(X \leqslant t) = \int_{-\infty}^{t} f_X(x)\,dx$$

CDF of uniform distribution

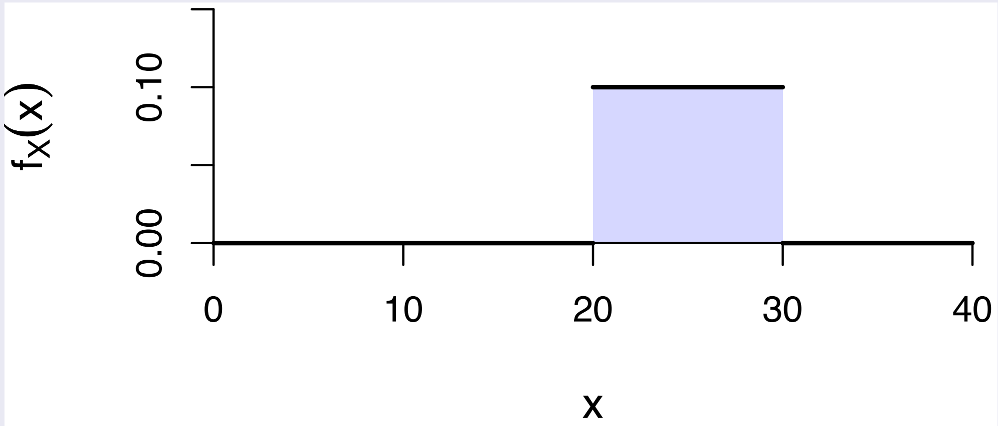
Uniform distribution on $[20, 30]$

- For $x < 20$: $F_X(x) = \int_{-\infty}^{x} 0\,dt = 0$
- For $20 \leqslant x < 30$: $F_X(x) = \int_{-\infty}^{20} 0\,dt + \int_{20}^{x} \frac{1}{10} dt = \frac{x-20}{10}$
- For $30 \leqslant x$: $F_X(x) = \int_{-\infty}^{20} 0\,dt + \int_{20}^{30} \frac{1}{10}\,dt + \int_{30}^{x} 0\,dt = 1$
- Together:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 20 \\ \frac{x-20}{10} & \text{if } 20 \leqslant x \leqslant 30 \\ 1 & \text{if } x \geqslant 30 \end{cases} \qquad f_X(x) = {F_X}'(x) = \begin{cases} 0 & \text{if } x < 20 \\ \frac{1}{10} & \text{if } 20 \leqslant x \leqslant 30 \\ 0 & \text{if } x \geqslant 30 \end{cases}$$

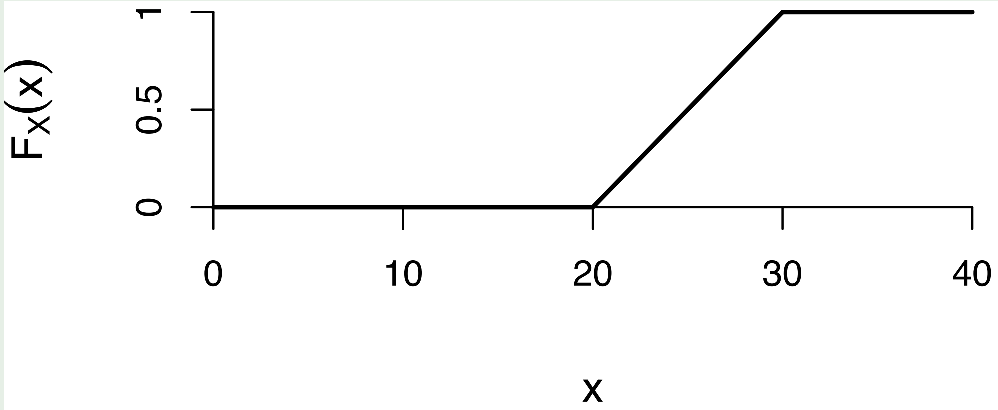
PDF vs. CDF

Probability density function



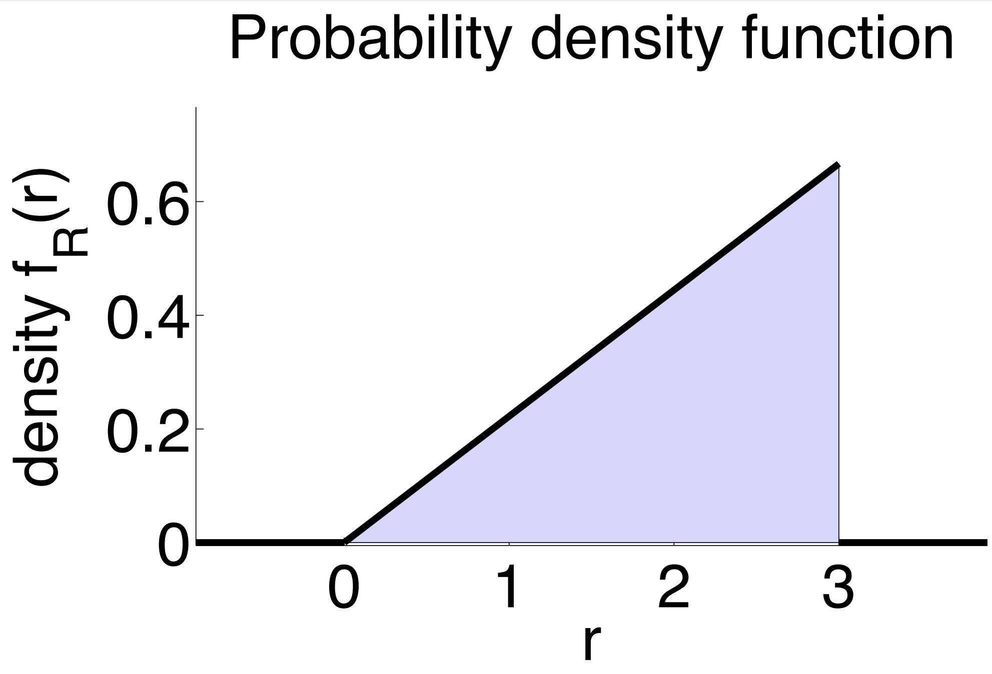
- $f_X(x) = \begin{cases} .1 & \text{if } 20 \leqslant x \leqslant 30; \\ 0 & \text{otherwise.} \end{cases}$
  It's discontinuous at $x = 20$ and 30.
- **PDF is derivative of CDF:**
  $f_X(x) = F_X{}'(x)$

Cumulative distribution function

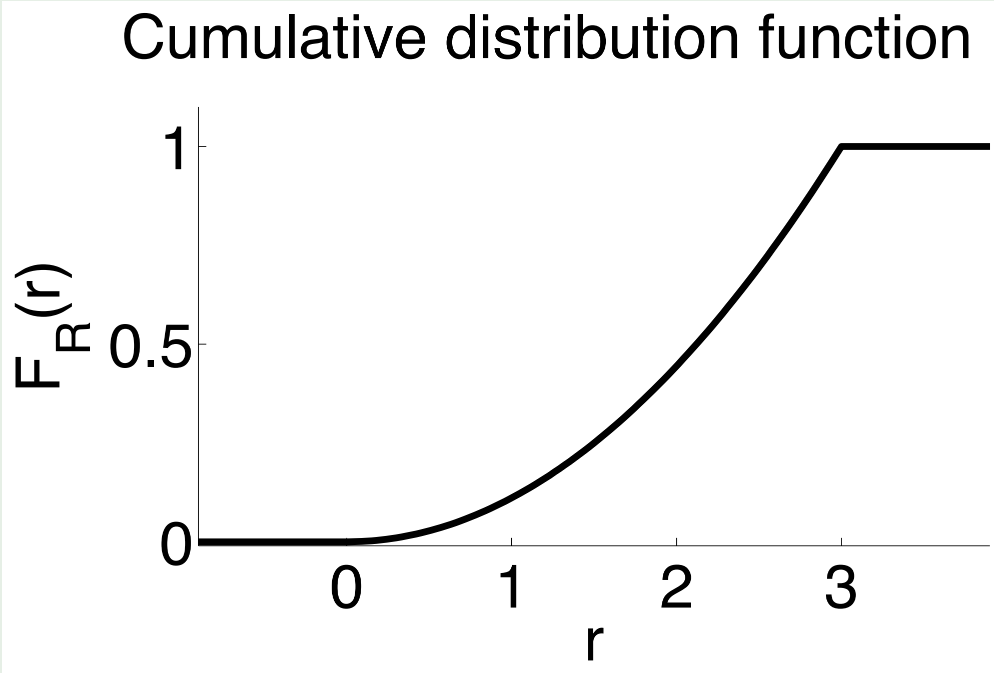


- $F_X(x) =$
  $\begin{cases} 0 & \text{if } x < 20; \\ (x-20)/10 & \text{if } 20 \leqslant x \leqslant 30; \\ 1 & \text{if } x \geqslant 30. \end{cases}$
- **CDF is integral of PDF:**
  $$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

PDF vs. CDF: Second example



- $f_R(r) = \begin{cases} 2r/9 & \text{if } 0 \leqslant r < 3; \\ 0 & \text{if } r \leqslant 0 \text{ or } r > 3 \end{cases}$
  It's discontinuous at $r = 3$.
- **PDF is derivative of CDF:**
  $f_R(r) = {F_R}'(r)$



- $F_R(r) = \begin{cases} 0 & \text{if } r < 0; \\ r^2/9 & \text{if } 0 \leqslant r \leqslant 3; \\ 1 & \text{if } r \geqslant 3. \end{cases}$
- **CDF is integral of PDF:**
  $F_R(r) = \int_{-\infty}^{r} f_R(t)\, dt$

Probability of an interval

Compute $P(-1 \leqslant R \leqslant 2)$ from the PDF and also from the CDF

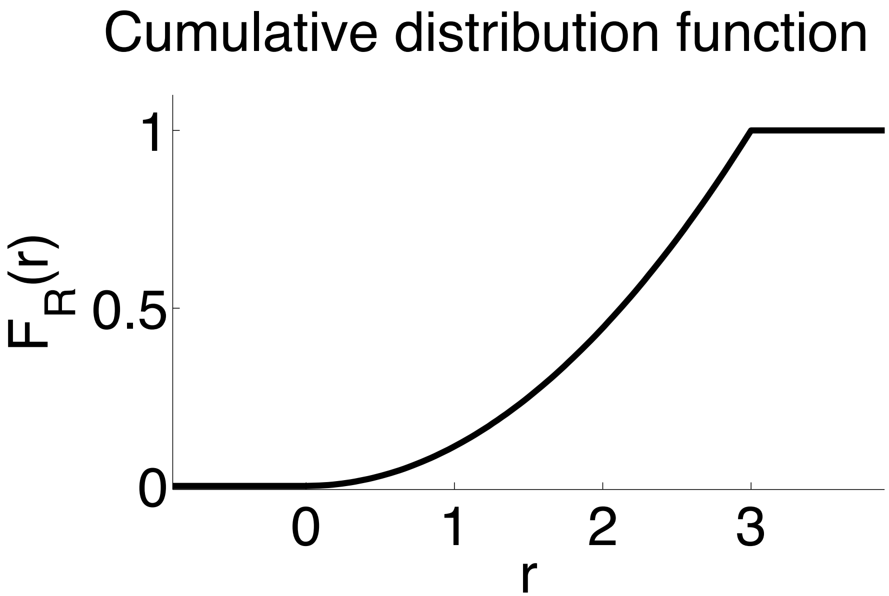
Computation from the PDF

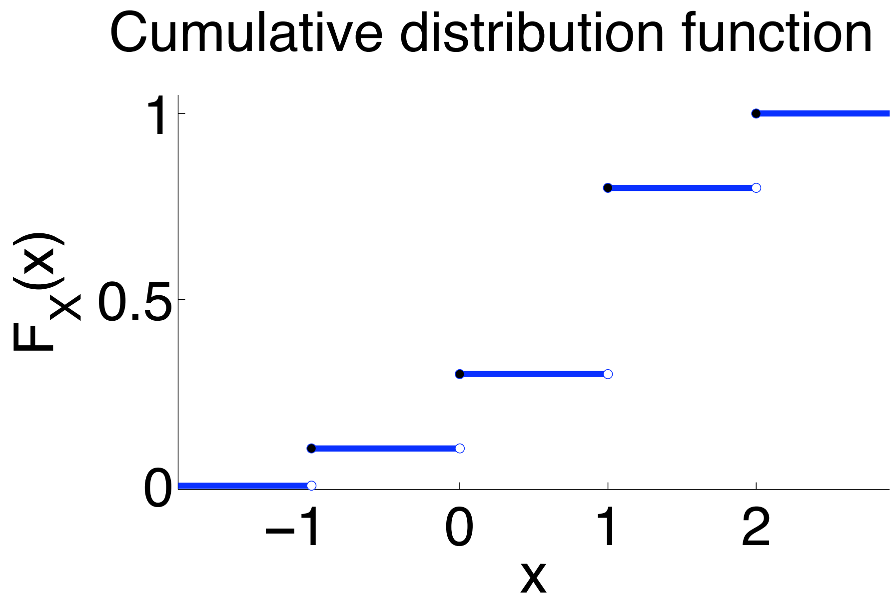
$$P(-1 \leqslant R \leqslant 2) = \int_{-1}^{2} f_R(r)\,dr = \int_{-1}^{0} f_R(r)\,dr + \int_{0}^{2} f_R(r)\,dr$$
$$= \int_{-1}^{0} 0\,dr + \int_{0}^{2} \frac{2r}{9}\,dr$$
$$= 0 + \left(\left. \frac{r^2}{9} \right|_{r=0}^{2} \right) = \frac{2^2 - 0^2}{9} = \boxed{\mathbf{\frac{4}{9}}}$$

Computation from the CDF

$$P(-1 \leqslant R \leqslant 2) = P(-1^- < R \leqslant 2)$$
$$= F_R(2) - F_R(-1^-) = \frac{2^2}{9} - 0 = \boxed{\mathbf{\frac{4}{9}}}$$

Continuous vs. discrete random variables





In a continuous distribution:

- The probability of an individual point is 0: $P(R = r) = 0$.
  So, $P(R \leqslant r) = P(R < r)$, i.e., $F_R(r) = F_R(r^-)$.
- The CDF is continuous.
  (In a discrete distribution, the CDF is discontinuous due to jumps at the points with nonzero probability.)
- $$P(a < R < b) = P(a \leqslant R < b) = P(a < R \leqslant b) = P(a \leqslant R \leqslant b)$$
  $$= F_R(b) - F_R(a)$$

Cumulative distribution function (CDF)

The *Cumulative Distribution Function (CDF)* of a random variable X is

$$F_X(x) = P(X \leqslant x)$$

Continuous case

- $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
- Weakly increasing.
- Varies smoothly from 0 to 1 as x varies from $-\infty$ to ∞.
- To get the PDF from the CDF, use $f_X(x) = {F_X}'(x)$.

Discrete case

- $F_X(x) = \sum_{t \leqslant x} P_X(t)$
- Weakly increasing.
- Stair-steps from 0 to 1 as x goes from $-\infty$ to ∞.
- The CDF jumps where $P_X(x) \neq 0$ and is constant in-between.
- To get the PDF from the CDF, use $P_X(x) = F_X(x) - F_X(x^-)$ (which is positive at the jumps, 0 otherwise).

CDF, percentiles, and median

The k^{th} *percentile* of a distribution X is the point x where k% of the probability is up to that point:

$$F_X(x) = P(X \leqslant x) = k\% = k/100$$

Example: $F_R(r) = P(R \leqslant r) = r^2/9 \quad$ (for $0 \leqslant r \leqslant 3$)

- $r^2/9 = (k/100) \quad \Rightarrow \quad r = \sqrt{9(k/100)}$
- 75^{th} percentile: $r = \sqrt{9(.75)} \approx 2.60$
- Median (50^{th} percentile): $r = \sqrt{9(.50)} \approx 2.12$
- 0^{th} and 100^{th} percentiles:
  $r = 0$ and $r = 3$ if we restrict to the range $0 \leqslant r \leqslant 3$.

  But they are not uniquely defined, since
  $F_R(r) = 0$ for all $r \leqslant 0$ and $F_R(r) = 1$ for all $r \geqslant 3$.

Expected value and variance (continuous r.v.)

Replace sums by integrals. It's the same definitions in terms of "$E(\cdot)$":

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X) \\ &= E((X-\mu)^2) = E(X^2) - (E(X))^2\end{aligned}$$

μ and σ for the uniform distribution on $[a, b]$ (with $a < b$)

$$\mu = E(X) = \int_a^b x \cdot \frac{1}{b-a}\,dx = \left.\frac{x^2/2}{b-a}\right|_{x=a}^{b} = \frac{(b^2-a^2)/2}{b-a} = \frac{b+a}{2}$$

$$E(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \left.\frac{x^3/3}{b-a}\right|_{x=a}^{b} = \frac{(b^3-a^3)/3}{b-a} = \frac{b^2+ab+a^2}{3}$$

$$\sigma^2 = \mathrm{Var}(X) = E(X^2) - (E(X))^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}$$

$$\sigma = \mathrm{SD}(X) = (b-a)/\sqrt{12}$$

Exponential distribution

- How far is it from the start of a chromosome to the first crossover?
- How far is it from one crossover to the next?
- Let D be the random variable giving either of those. It is a real number > 0, with the *exponential distribution*

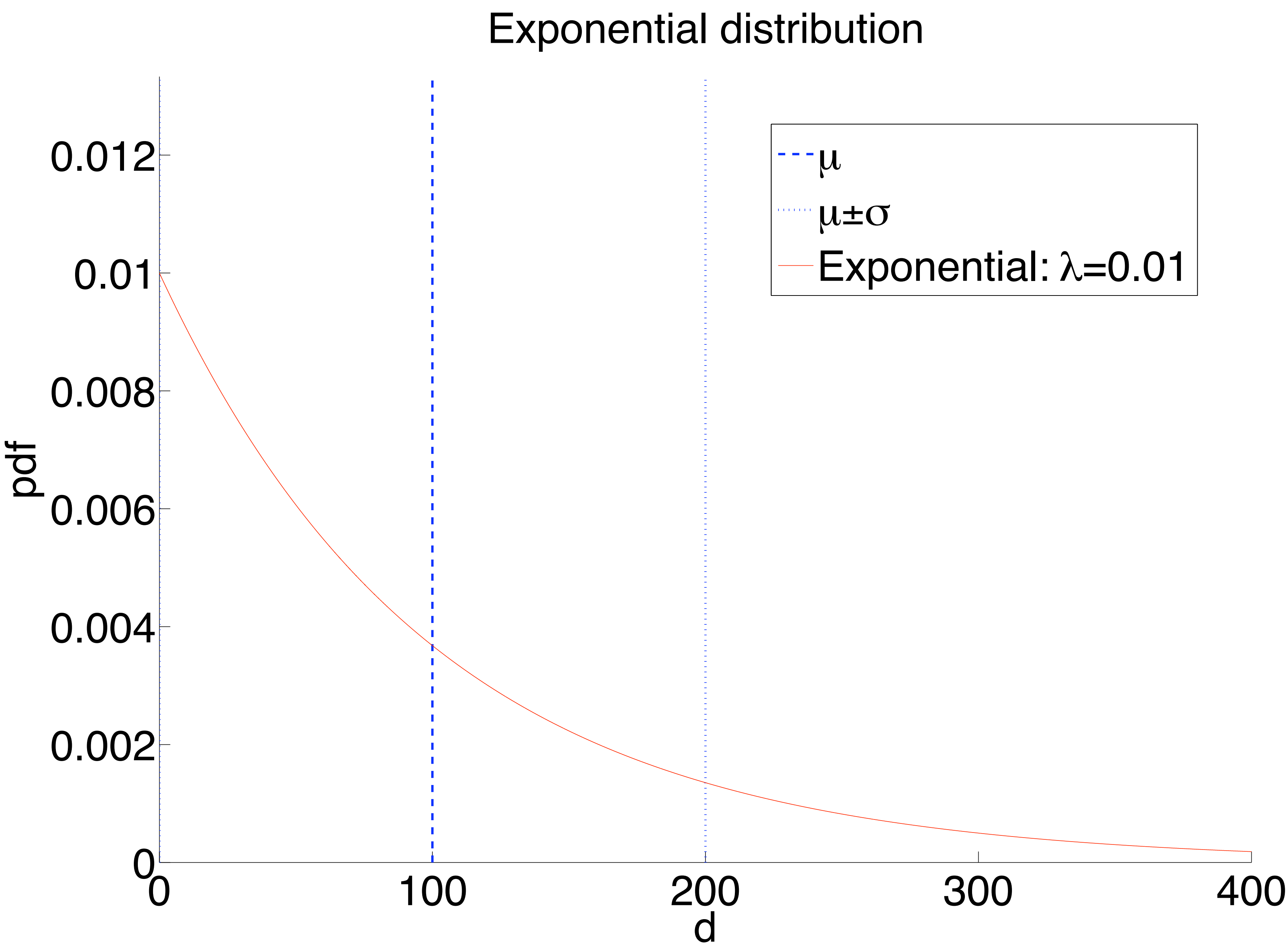
$$f_D(d) = \begin{cases} \lambda\, e^{-\lambda\, d} & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

  where crossovers happen at a rate $\lambda = 1$ M^{-1} $= 0.01$ cM^{-1}.

-

	General case		**Crossovers**
Mean	$E(D) = 1/\lambda$	$=$	100 cM $= 1$ M
Variance	$\mathrm{Var}(D) = 1/\lambda^2$	$=$	10000 cM^2 $= 1$ M^2
Standard Dev.	$\mathrm{SD}(D) = 1/\lambda$	$=$	100 cM $= 1$ M

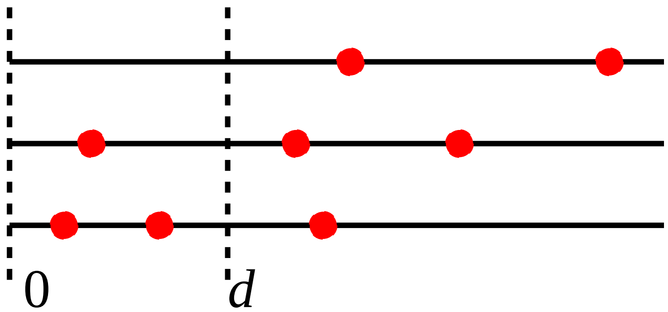
Exponential distribution



Exponential distribution

- In general, if events occur on the real number line $x \geqslant 0$ in such a way that the expected number of events in all intervals $[x, x+d]$ is λd (for $x > 0$), then the exponential distribution with parameter λ models the time/distance/etc. until the first event.
- It also models the time/distance/etc. between consecutive events.
- Chromosomes are finite; to make this model work, treat "there is no next crossover" as though there is one but it happens somewhere past the end of the chromosome.

Proof of PDF formula for exponential distribution



- Let $d > 0$ be any positive real number.
- Let $N(d)$ be the # of crossovers that occur in the interval $[0, d]$.
  - If $N(d) = 0$ then there are no crossovers in $[0, d]$, so $D > d$.
  - If $D > d$ then the first crossover is after d so $N(d) = 0$.
  - Thus, $D > d$ is equivalent to $N(d) = 0$
    and $D \leqslant d$ is equivalent to $N(d) > 0$.
- $P(D > d) = P(N(d) = 0) = e^{-\lambda d}(\lambda d)^0/0! = e^{-\lambda d}$
  since $N(d)$ has a Poisson distribution with parameter λd.

Proof of PDF formula for exponential distribution

$$P(D > d) = \begin{cases} e^{-\lambda d} & \text{if } d \geqslant 0 \text{ (from previous slide);} \\ 1 & \text{if } d < 0 \\ & (D \text{ is positive, so } D > \text{ any negative number)} \end{cases}$$

CDF of D

$$F_D(d) = P(D \leqslant d) = 1 - P(D > d) = \begin{cases} 1 - e^{-\lambda d} & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

Differentiate CDF and simplify to get PDF

$$f_D(d) = \begin{cases} \lambda e^{-\lambda d} & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

Discrete and Continuous Analogs

	Discrete	**Continuous**
"Success"	Coin flip at a position is heads	Point where crossover occurs
Rate	Probability p per flip	λ (crossovers per Morgan)
# successes	Binomial distribution: # heads out of n flips	Poisson distribution: # crossovers in distance d
Wait until 1st success	Geometric distribution	Exponential distribution
Wait until r^{th} success	Negative binomial distribution	Gamma distribution

Gamma distribution

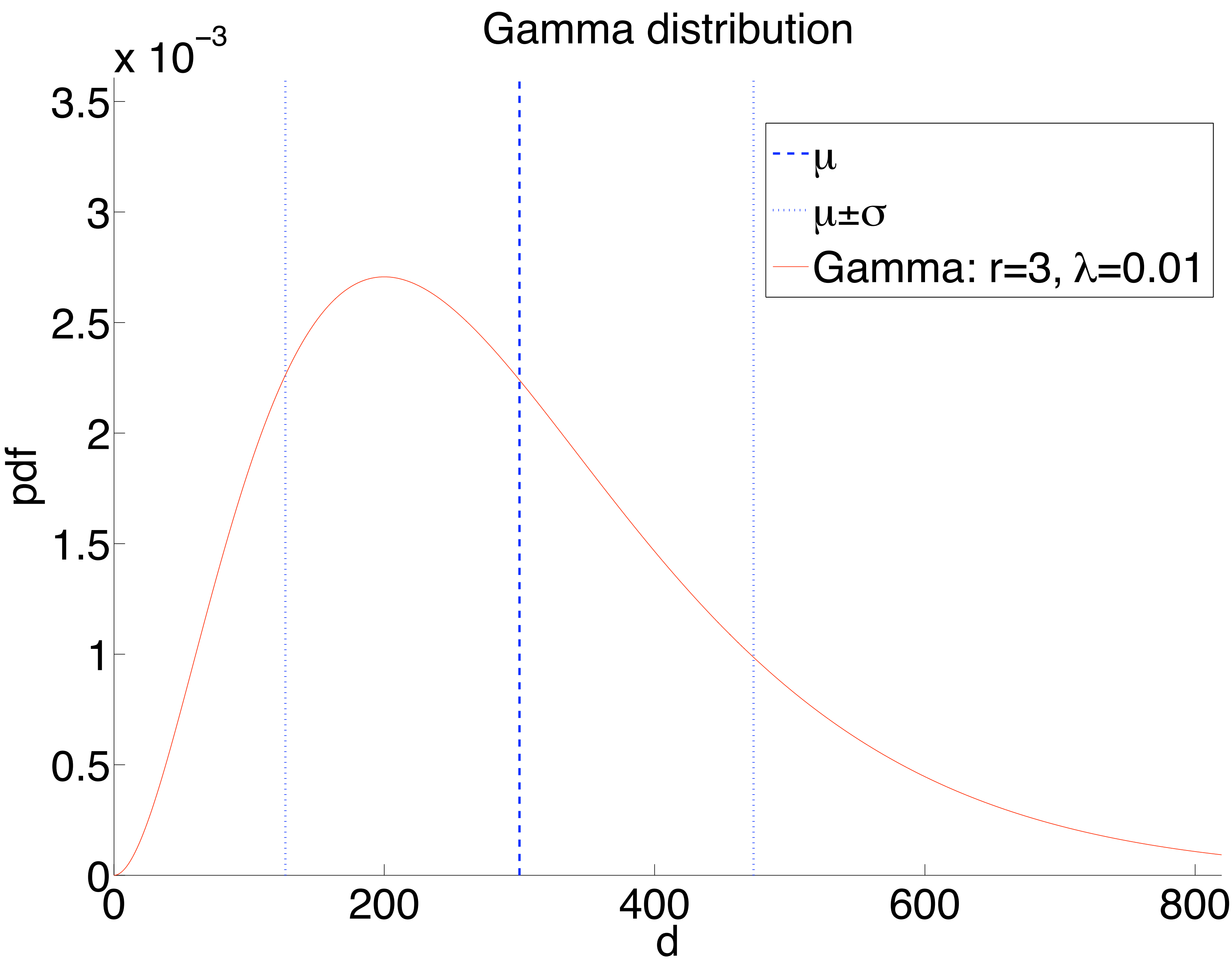
- How far is it from the start of a chromosome until the r^{th} crossover, for some choice of $r = 1, 2, 3, \ldots$?
- Let D_r be a random variable giving this distance.
- It has the *gamma distribution* with PDF

$$f_{D_r}(d) = \begin{cases} \frac{\lambda^r}{(r-1)!} d^{r-1} e^{-\lambda d} & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

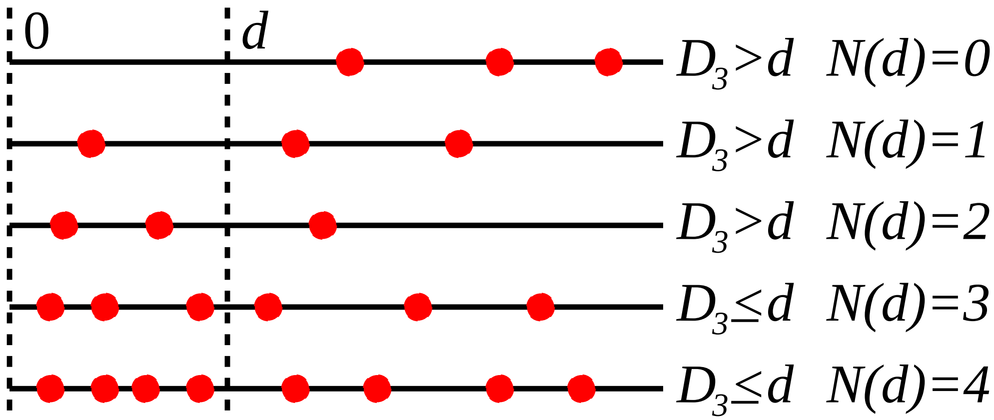
Mean	$E(D_r) = r/\lambda$
Variance	$\text{Var}(D_r) = r/\lambda^2$
Standard deviation	$\text{SD}(D_r) = \sqrt{r}/\lambda$

- The gamma distribution for $r = 1$ is the same as the exponential distribution.
- The sum of r i.i.d. exponential variables, $D_r = X_1 + X_2 + \cdots + X_r$, each with rate λ, gives the gamma distribution.

Gamma distribution



Proof of Gamma distribution PDF for $r = 3$



- Let $d > 0$ be any real number.
- $D_3 > d$ is the event the 3rd crossover is after position d.
  - Then the number of crossovers in $[0, d]$ is < 3, so $N(d) < 3$.
  - $D_3 > d$ is equivalent to $N(d) < 3$.
  - $D_3 \leqslant d$ is equivalent to $N(d) \geqslant 3$.

Proof of Gamma distribution PDF for $r = 3$

- $D_3 > d$ is the event the 3rd crossover is after position d.
  It's equivalent to $N(d) < 3$, so $N(d)$ is 0, 1, or 2:

$$\begin{aligned} P(D_3 > d) &= P(N(d)=0) + P(N(d)=1) + P(N(d)=2) \\ &= e^{-\lambda d}\left(\frac{(\lambda d)^0}{0!} + \frac{(\lambda d)^1}{1!} + \frac{(\lambda d)^2}{2!}\right) \end{aligned}$$

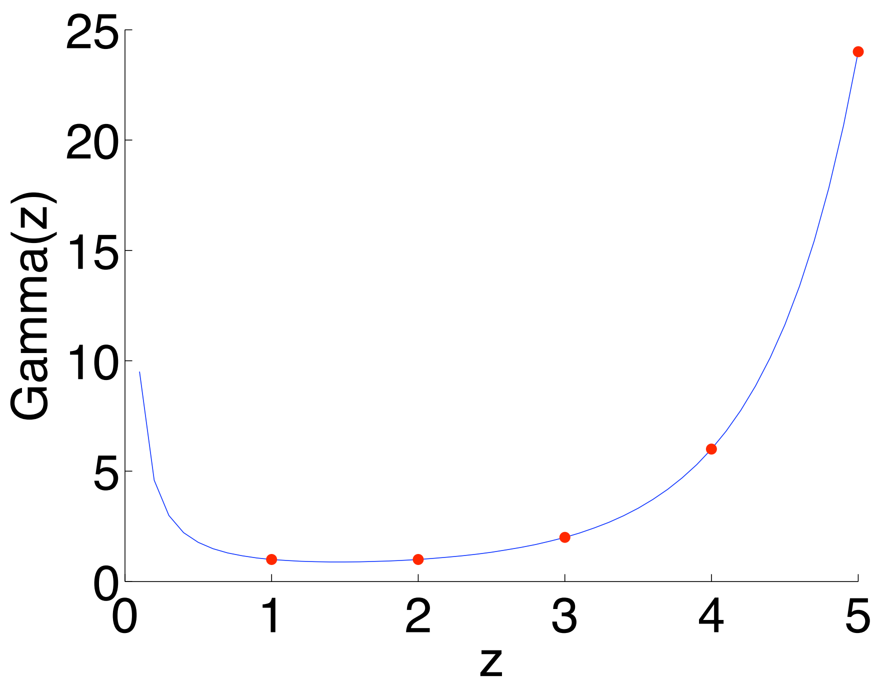
- The CDF of D_3 is $P(D_3 \leqslant d) = 1 - P(D_3 > d)$.

- Differentiating the CDF and simplifying gives the PDF

$$f_{D_3}(d) = \begin{cases} \lambda^3 d^2 e^{-\lambda d}/2! & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

The Gamma function and factorials

- The *Gamma function* is a generalization of factorials:
$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$$
for real $z > 0$.
- $\Gamma(z) = (z-1)!$ for $z = 1, 2, 3, \ldots$
- $\Gamma(z)$ extends to all complex numbers except integers $\leqslant 0$.



Proof of $\Gamma(z) = (z-1)!$ for $z = 1, 2, 3, \ldots$.

- $\Gamma(1) = \int_0^\infty t^0 e^{-t}\, dt = -e^{-t}\Big|_0^\infty = -0 + 1 = 1$
- $\Gamma(z) = (z-1)\Gamma(z-1)$ can be shown using integration by parts: differentiate t^{z-1} and integrate up $e^{-t}\, dt$.
- When z is a positive integer, iterate this to
$\Gamma(z) = (z-1)(z-2)\cdots(2)(1)\Gamma(1) = (z-1)! \cdot \Gamma(1) = (z-1)!$ $\square$

Variations of the distributions

- The Gamma distribution is defined for real $r > 0$ rather than just positive integers, by replacing $(r-1)!$ with $\Gamma(r)$:

$$f_{D_r}(d) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} d^{r-1} e^{-\lambda d} & \text{if } d \geqslant 0; \\ 0 & \text{if } d < 0. \end{cases}$$

- **Upcoming:** Chi-squared distribution has $r = n/2$ (half-integers).

- For Poisson, Exponential, and Gamma distributions, instead of the rate parameter λ, some people use the *shape* parameter $\theta = 1/\lambda$:
  - For crossovers, $\theta = 1$ M $= 100$ cM.
  - The Poisson parameter for distance d is $\mu = \lambda d = d/\theta$.