

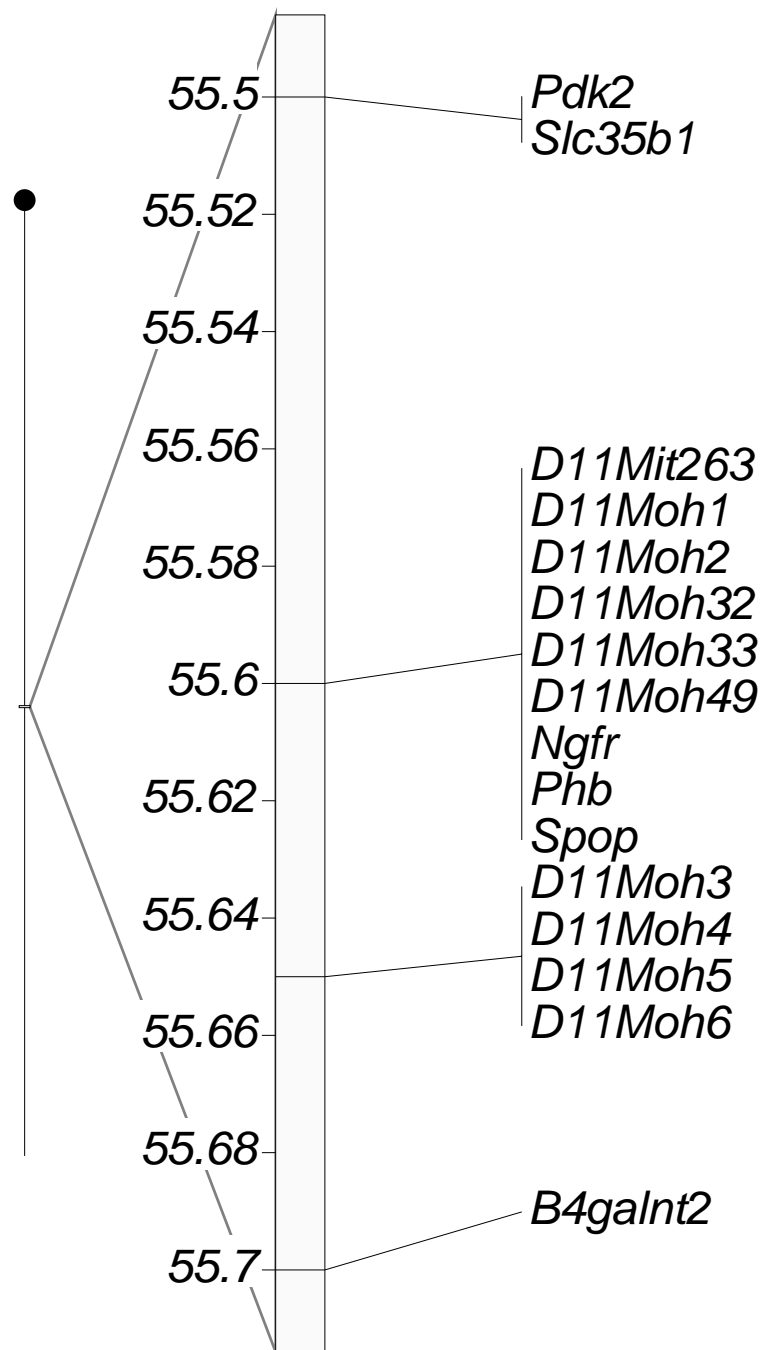
4.2 Poisson Distribution: Counting Crossovers in Meiosis

4.2 Exponential and 4.6 Gamma Distributions: Distance Between Crossovers

Prof. Tesler

Math 186
February 18, 2009

Chromosomal coordinate system



- **Mouse chr. 11: 55.50–55.70 cM. Linkage map obtained Feb. 17, 2008 from <http://www.informatics.jax.org>**
- The unit *Morgan* is defined so that crossovers occur at an average rate 1 per Morgan (M) or .01 per centi-Morgan (cM).
- 1911–1913: Alfred H. Sturtevant developed the first genetic map of a chromosome, *D. melanogaster* (fruit fly), as an undergrad in Thomas Hunt Morgan's lab.
- 1919: J.B.S. Haldane improved on it and renamed the units to Morgans and centi-Morgans.

Morgans and centi-Morgans

- Morgans (M) and centi-Morgans ($1 \text{ cM} = .01 \text{ M}$) are a coordinate system for chromosomes based on recombination rates during meiosis.
- They are expressed as a real number ≥ 0 .
- Two genes on the same chromosome at positions d_1 and d_2 in Morgans, have an average of $|d_1 - d_2|$ crossovers between them.
- It's more common to measure it in centi-Morgans, so two genes located 123 cM apart would have an average of 1.23 crossovers between them.
- *Units of (centi-)Morgans were developed prior to the discovery that DNA is comprised of a large but finite number of discrete nucleotides.*

Counting crossovers

- *Assumption 1:* Crossovers between two genes on the same chromosome occur at a rate

$$\begin{aligned}\lambda &= 0.01 \text{ per cM} = 0.01 \text{ cM}^{-1} \\ &= 1 \text{ per M} = 1 \text{ M}^{-1}\end{aligned}$$



If genes A and B are $d = 123$ cM apart, the average number of crossovers between them per meiosis over the whole population is

$$\lambda d = (0.01 \text{ cM}^{-1})(123 \text{ cM}) = 1.23$$

- $\lambda d = 1.23 > 0$ is a pure number without units.
- *Assumption 2:* If genes are in order A, B, C , then crossovers between A and B are independent of crossovers between B and C .
- **What is the probability that k crossovers occur between A and B ?**

Counting crossovers



- Let $X = 0, 1, 2, \dots$ be the number of crossovers occurring between A and B in a particular meiosis.
- X is a discrete random variable. We will develop a discrete pdf for it called the *Poisson distribution*.
- We'll use the average number of crossovers between them (e.g., $\lambda d = 1.23$) to determine the distribution of X .

Counting crossovers with the binomial distribution



Split the interval from A to B into n “equal” pieces and assume that:

- The probability of 2 or more crossovers in a piece is essentially 0. (*This requires n to be large.*)
- Crossovers in different pieces occur independently.
- Crossover probabilities are the same in each piece. (*This is what makes the pieces “equal.”*)

In this model, the number of crossovers, X , follows a binomial distribution:

- There are n pieces, each with (unknown) equal probability p of having a crossover, so the average number of crossovers is np .
- The average is also λd , so $np = \lambda d$ and $p = \lambda d/n$.
- For $k = 0, 1, 2, \dots$

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} = \binom{n}{k} \left(\frac{\lambda d}{n}\right)^k \left(1 - \frac{\lambda d}{n}\right)^{n-k}$$

Counting crossovers with the binomial distribution

- Suppose $d = 123$ cM (so $\lambda d = 1.23$) and $k = 3$.
- We don't know n ; however, as $n \rightarrow \infty$, the digits of $P(X = 3)$ stabilize:

n	$p = \lambda d/n$	Binomial pdf $P(X = 3) = \binom{n}{3} p^3 (1 - p)^{n-3}$
1	1.23	0
10^1	$1.23 \cdot 10^{-1}$	0.08910328876
10^2	$1.23 \cdot 10^{-2}$	0.09058485007
10^3	$1.23 \cdot 10^{-3}$	0.09064683438
10^4	$1.23 \cdot 10^{-4}$	0.09065233222
10^5	$1.23 \cdot 10^{-5}$	0.09065287510
10^6	$1.23 \cdot 10^{-6}$	0.09065292933
10^7	$1.23 \cdot 10^{-7}$	0.09065293476
10^8	$1.23 \cdot 10^{-8}$	0.09065293534

Theorem (Poisson Limit)

$$\lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} = \frac{e^{-\mu} \mu^k}{k!}$$

- For $\mu = \lambda d = 1.23$ and $k = 3$, this gives

$$\frac{e^{-1.23} 1.23^3}{3!} \approx 0.09065293537$$

- The values in the table converge to this as $n \rightarrow \infty$.
- Even for $n = 10$, the value in the table was within 2% of this limit.

Poisson limit – Proof for $k = 3$

$$\begin{aligned} \binom{n}{3} \left(\frac{\mu}{n}\right)^3 \left(1 - \frac{\mu}{n}\right)^{n-3} &= \frac{n(n-1)(n-2)}{3!} \frac{\mu^3}{n^3} \frac{\left(1 - \frac{\mu}{n}\right)^n}{\left(1 - \frac{\mu}{n}\right)^3} \\ &= \frac{\mu^3}{3!} \frac{n(n-1)(n-2)}{n^3} \frac{\left(1 - \frac{\mu}{n}\right)^n}{\left(1 - \frac{\mu}{n}\right)^3} \end{aligned}$$

Limits of each piece:

- $n(n-1)(n-2)/n^3 \rightarrow 1$
- $\left(1 - \frac{\mu}{n}\right)^3 \rightarrow (1-0)^3 = 1$
- $\left(1 - \frac{\mu}{n}\right)^n \rightarrow 1^\infty$; need L'Hospital's rule!
- L'Hospital's rule gives $\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x$, so $\lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}$

$$\binom{n}{3} \left(\frac{\mu}{n}\right)^3 \left(1 - \frac{\mu}{n}\right)^{n-3} \rightarrow \frac{\mu^3}{3!} \cdot 1 \cdot \frac{e^{-\mu}}{1} = \frac{\mu^3}{3!} e^{-\mu} \quad \square$$

Application

Historically, people approximated the binomial distribution by

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{e^{-np} (np)^k}{k!}$$

for $n \geq 50$ and $np \leq 5$.

Due to modern calculators and computers, this is not used as much.

Poisson distribution

- The second method to count crossovers is to define a new distribution based on the limiting process we just studied.
- The Poisson distribution with parameter μ (a positive real #) is

$$P(X = k) = \begin{cases} \frac{e^{-\mu} \mu^k}{k!} & \text{for } k = 0, 1, 2, \dots; \\ 0 & \text{otherwise.} \end{cases}$$

- It's a valid pdf since it's always ≥ 0 and the total probability is 1:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{e^{-\mu} \mu^k}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1$$

Poisson distribution – rates

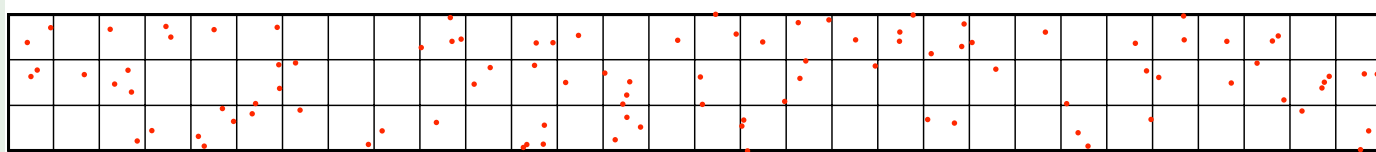
Poisson processes often involve rates, and the parameter may be described in terms of a rate:

Crossovers

- $\lambda = .01 \text{ cM}^{-1} = 1 \text{ M}^{-1}$ is called the *rate of a Poisson process*.
- $\mu = \lambda d$ is the *Poisson parameter*. It is a unitless number.

Other rates

- λ could be the average number of events per unit time, length, area, volume, etc., giving $\mu = \lambda t$, $\mu = \lambda \ell$, $\mu = \lambda A$, $\mu = \lambda V$, etc.
- E.g., collect rain on a rectangular area for 1 second, and let λ be the average number of raindrops per unit area per second:



Then for area A and time t , the expected number of raindrops is $\mu = \lambda A t$.

Probabilities of different numbers of crossovers

- This table shows the probability of k crossovers occurring during meiosis, for two genes located 123 cM apart ($\mu = 1.23$).
- If we look at 100 gametes formed independently (say in different individuals), the expected # exhibiting k crossovers is $100 P(X = k)$.

# crossovers k	Theoretical proportion (pdf) $P(X = k) = e^{-1.23} (1.23)^k / k!$	Frequency $100 P(X = k)$
0	.2922925777	29.2292577
1	.3595198706	35.95198706
2	.2211047204	22.11047204
3	.09065293537	9.065293537
4	.02787577763	2.787577763
5	.006857441295	0.6857441295
6	.001405775465	0.1405775465
7	.0002470148317	0.02470148317
...

- $P(X = 1.5) = 0$, $P(X = -2) = 0$ (not non-negative integers)
- $P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) = 0.1270828313$
- Theoretical frequency of $X \geq 3$ is $100P(X \geq 3) = 12.70828313$.

Mean and standard deviation of Poisson Distribution

- The mean of the Poisson distribution equals the Poisson parameter (which is why we can call the parameter μ).
- The variance is $\sigma^2 = \mu$ and the standard deviation is $\sigma = \sqrt{\mu}$.
- In the $d = 123$ cM example,
 - the average number of crossovers between the two sites is $\mu = \lambda d = 1.23$;
 - the variance of that is $\sigma^2 = 1.23$;
 - the standard deviation is $\sigma = \sqrt{1.23} \approx 1.11$

Proof.

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \cdot \frac{e^{-\mu} \mu^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\mu} \mu^k}{(k-1)!} = \mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} \\ &= \mu e^{-\mu} e^{\mu} = \mu \end{aligned}$$

That proves it for the mean.

$E(X^2) = \mu(\mu + 1)$ can be shown in a similar fashion, so the variance is

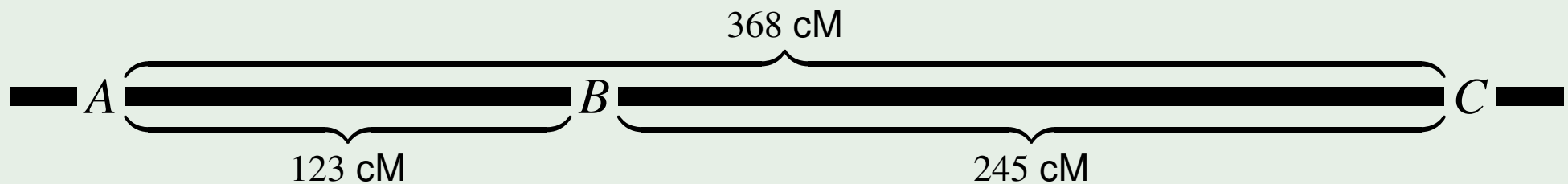
$$\text{Var}(X) = E(X^2) - (E(X))^2 = \mu(\mu + 1) - \mu^2 = \mu. \quad \square$$

Sum of Poisson Random Variables

Theorem (Sum of Poisson Random Variables)

Let X, Y be independent Poisson random variables with parameters μ and ν . Then $W = X + Y$ is Poisson with parameter $\mu + \nu$.

Example



The # crossovers between A & B is Poisson with parameter 1.23

B & C	2.45
A & C	<u>3.68</u>

Addition of Poisson Random Variables

Theorem (Addition of Poisson Random Variables)

Let X, Y be independent Poisson random variables with parameters μ and ν . Then $W = X + Y$ is Poisson with parameter $\mu + \nu$.

Proof.

$$\begin{aligned}P(W = k) &= \sum_{m=0}^k P(X = m)P(Y = k - m) \\&= \sum_{m=0}^k \frac{e^{-\mu} \mu^m}{m!} \cdot \frac{e^{-\nu} \nu^{k-m}}{(k-m)!} = e^{-(\mu+\nu)} \sum_{m=0}^k \frac{\mu^m \nu^{k-m}}{m! (k-m)!} \\&= \frac{e^{-(\mu+\nu)}}{k!} \sum_{m=0}^k \frac{k!}{m! (k-m)!} \mu^m \nu^{k-m} \\&= \frac{e^{-(\mu+\nu)}}{k!} \sum_{m=0}^k \binom{k}{m} \mu^m \nu^{k-m} = \frac{e^{-(\mu+\nu)} (\mu + \nu)^k}{k!} \quad \square\end{aligned}$$

Determining the Poisson parameter from data

Suppose that we had a way to count many crossovers occurred between two genes in individual meioses, and we count it in 100 independent gametes as shown in the table below. How far apart are the genes in cM?

k	Obs. Freq.	Obs. Prop.	# Crossovers
0	64	0.64	0
1	29	0.29	29
2	6	0.06	12
3	1	0.01	3
Total	100	1.00	44

- **Observed Frequency:** # gametes with exactly k crossovers between A and B
- **Observed Proportion:** frequency / total # gametes
- **# Crossovers:** Total number of crossovers accounted for = k times observed frequency

Determining the Poisson parameter from data

k	Obs. Freq.	Obs. Prop.	# Crossovers
0	64	0.64	0
1	29	0.29	29
2	6	0.06	12
3	1	0.01	3
Total	100	1.00	44

- The total number of crossovers between A and B that occurred among all 100 gametes is $0(64) + 1(29) + 2(6) + 3(1) = 44$.
- The average number of crossovers per gamete is $44/100 = 0.44$.
- The Poisson parameter is $0.44 = \lambda d$ so they are about $d = 0.44/\lambda = 44$ cM apart.
- **Note:** these are easy numbers to demonstrate the general procedure for fitting the Poisson distribution, but are unrealistic for crossovers. Linkage maps are constructed using markers much closer together, $\ll 1$ cM, to make sure that only $k = 0$ and $k = 1$ arise.

Determining the Poisson parameter from data

Compare the original data with the values predicted by the Poisson distribution for $d = 44$ cM. Observed and theoretical values are close:

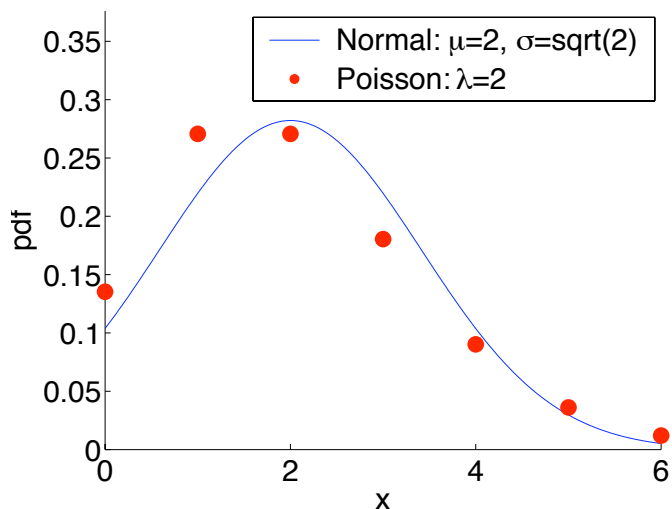
k	Obs. Freq.	Obs. Prop.	# Crossovers
0	64	0.64	0
1	29	0.29	29
2	6	0.06	12
3	1	0.01	3
Total	100	1.00	44

k	Theoretical proportion (pdf)	Theoretical frequency
	$P(X = k) = e^{-0.44} (0.44)^k / k!$	$100 P(X = k)$
0	.6440364211	64.40364211
1	.2833760253	28.33760253
2	.06234272555	6.234272555
3	.009143599749	.9143599749
4	.001005795973	.1005795973
5	.00008851004558	.008851004558
...

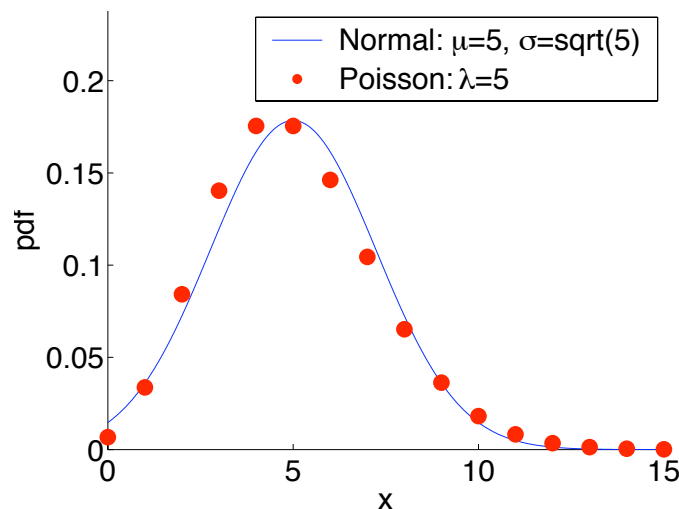
Poisson and normal distributions

When $\mu \geq 5$, the Poisson distribution is also well-approximated by the normal distribution with the same μ and with $\sigma = \sqrt{\mu}$:

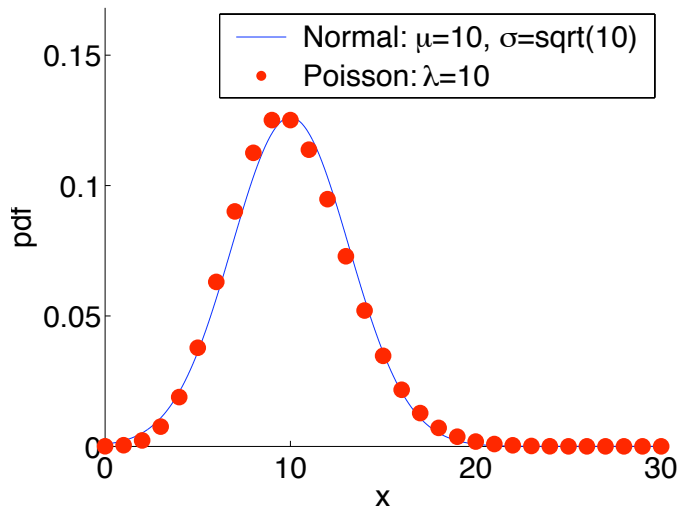
Comparison of normal and Poisson distributions



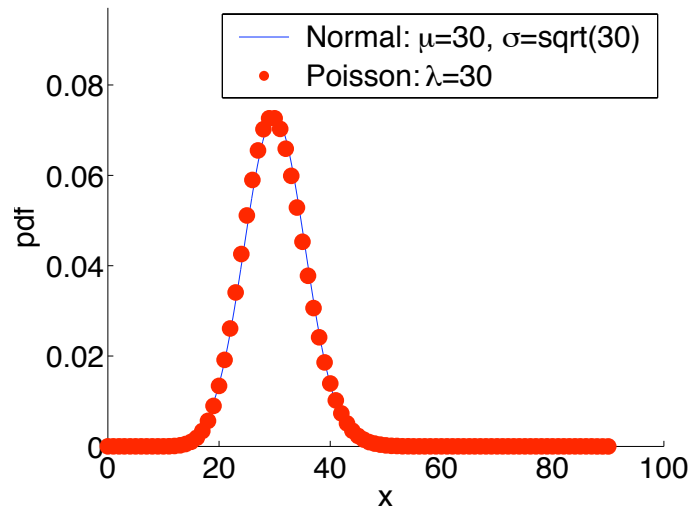
Comparison of normal and Poisson distributions



Comparison of normal and Poisson distributions



Comparison of normal and Poisson distributions



4.2 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*

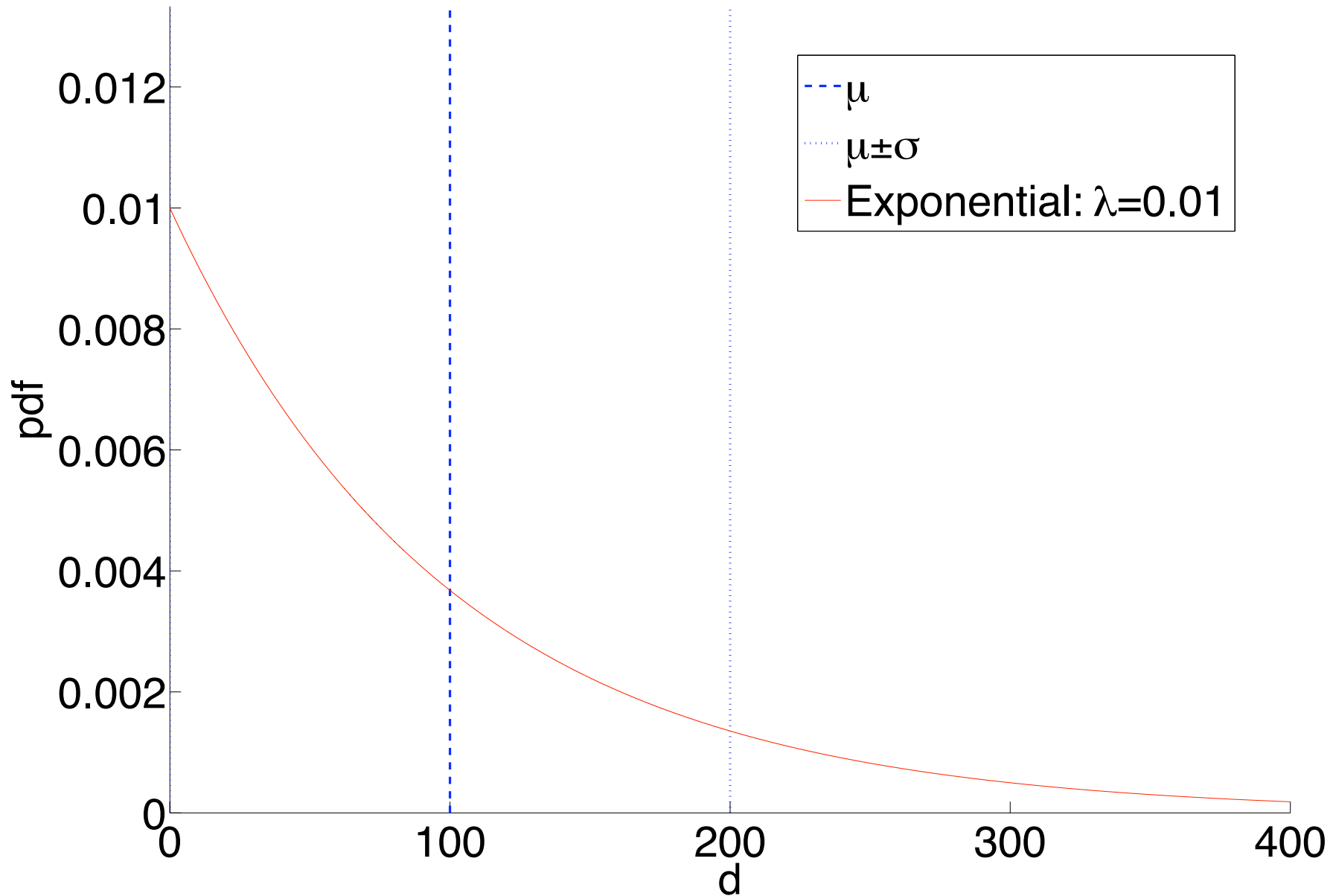
$$f_D(d) = \begin{cases} \lambda e^{-\lambda d} & \text{if } d \geq 0; \\ 0 & \text{if } d < 0. \end{cases}$$

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

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

4.2 Exponential distribution

Exponential distribution

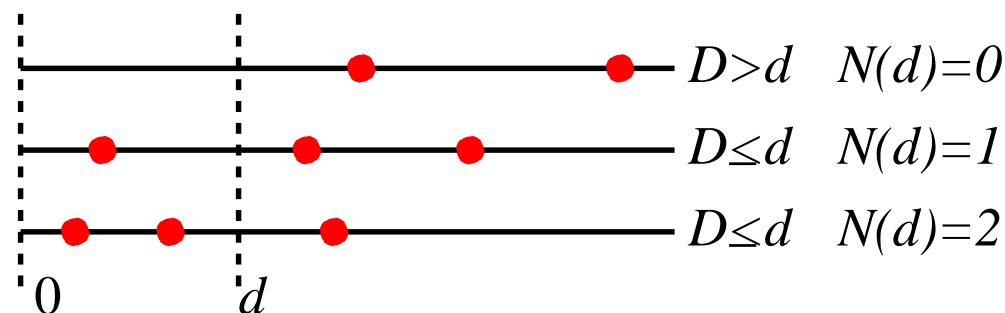


4.2 Exponential distribution

- In general, if events occur on the real number line $x \geq 0$ 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

- Let $d > 0$ be any 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 $N(d) = 0$.
- $P(D > d) = P(N(d) = 0) = e^{-\lambda d} (\lambda d)^0 / 0! = e^{-\lambda d}$
since $N(d)$ has a Poisson distribution with parameter λd .
- The cdf of D is
$$F_D(d) = P(D \leq d) = 1 - P(D > d) = \begin{cases} 1 - e^{-\lambda d} & \text{if } d \geq 0; \\ 0 & \text{if } d < 0. \end{cases}$$
- Differentiating the cdf gives pdf $f_D(d) = F_D'(d) = \lambda e^{-\lambda d}$ (if $d \geq 0$).

Discrete and Continuous Analogs

	Discrete	Continuous
“Success”	Coin flip at a position is heads	Point where crossover occurs
Rate	Probability p per flip	λd crossovers per length d
# successes	Binomial distribution # heads out of n flips	Poisson distribution # crossovers in distance λd
Wait until 1st success	Geometric distribution	Exponential distribution
Wait until r th success	Negative binomial distribution	Gamma distribution

4.6 Gamma distribution

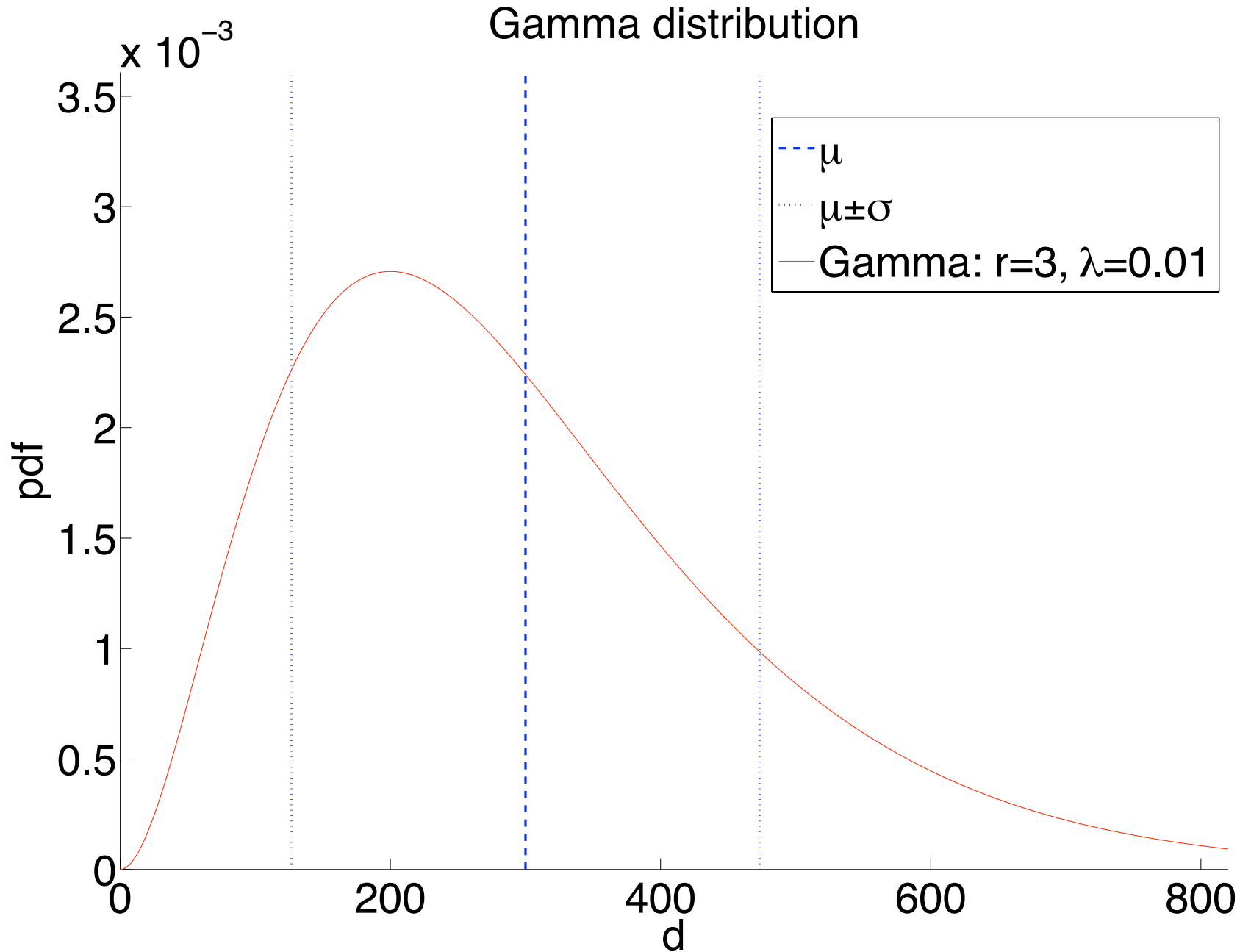
- How far is it from the start of a chromosome until the r th crossover, for some choice of $r = 1, 2, 3, \dots$?
- Let D_r be a random variable giving this distance.

- It has the *gamma distribution* with pdf

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

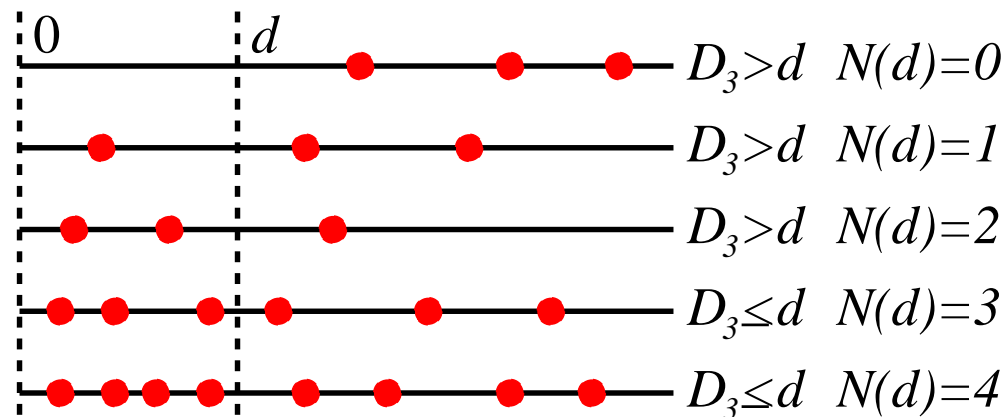
- **Mean** $E(D_r) = r/\lambda$
- **Variance** $\text{Var}(D_r) = r/\lambda^2$
- **Standard deviation** $\text{SD}(D_r) = \sqrt{r}/\lambda$
- The gamma distribution for $r = 1$ is the same as the exponential distribution.
- The sum of r i.i.d. exponential variables, $D_r = X_1 + X_2 + \dots + X_r$, each with rate λ , gives the gamma distribution.

4.6 Gamma distribution



Proof of Gamma distribution pdf for $r = 3$

- Let $d > 0$ be any real number.
- $D_3 > d$ is the event that the third crossover does not happen until sometime after position d .



- When $D_3 > d$, the number $N(d)$ of crossovers in the chromosome interval $[0, d]$ is less than 3, so it's 0, 1, or 2.
 $D_3 > d$ is equivalent to $N(d) < 3$.
 $D_3 \leq d$ is equivalent to $N(d) \geq 3$.

Proof of Gamma distribution pdf for $r = 3$

- Let $d > 0$ be any real number.
- $D_3 > d$ is the event that the third crossover does not happen until sometime after position d .
- When $D_3 > d$, the number $N(d)$ of crossovers in the chromosome interval $[0, d]$ is less than 3, so it's 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 $P(D_3 \leq d) = 1 - P(D_3 > d)$.
- Differentiating the cdf and simplifying gives the pdf

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