

Lander-Waterman Sequencing Statistics

Reference:

Lander ES, Waterman MS Genomic mapping by fingerprinting random clones: a mathematical analysis, Genomics 2(3): 231-239 (1988)

Poisson Distribution

Poisson distribution is used mostly to model rates in given intervals. For example, the number of accidents that occur in a given highway intersection in one week follows a Poisson distribution. The probability function is:

$P(Y=y) = (\lambda^y * e^{-\lambda}) / y!$ where $y = \#$ of events in a given interval,
 $\lambda =$ mean number of events in a given interval

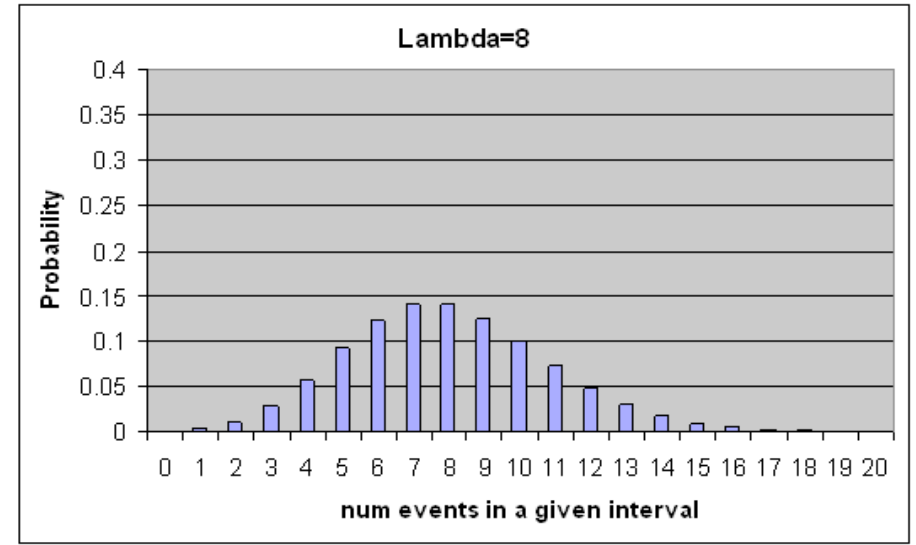
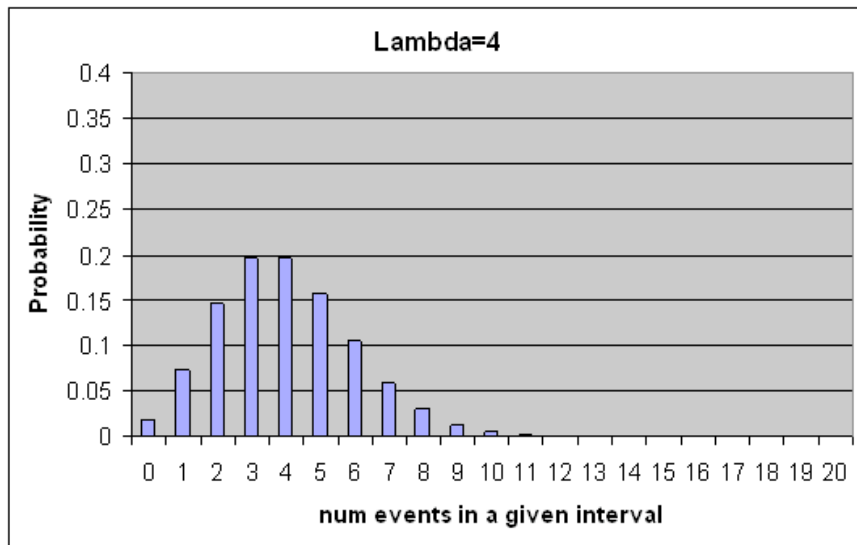
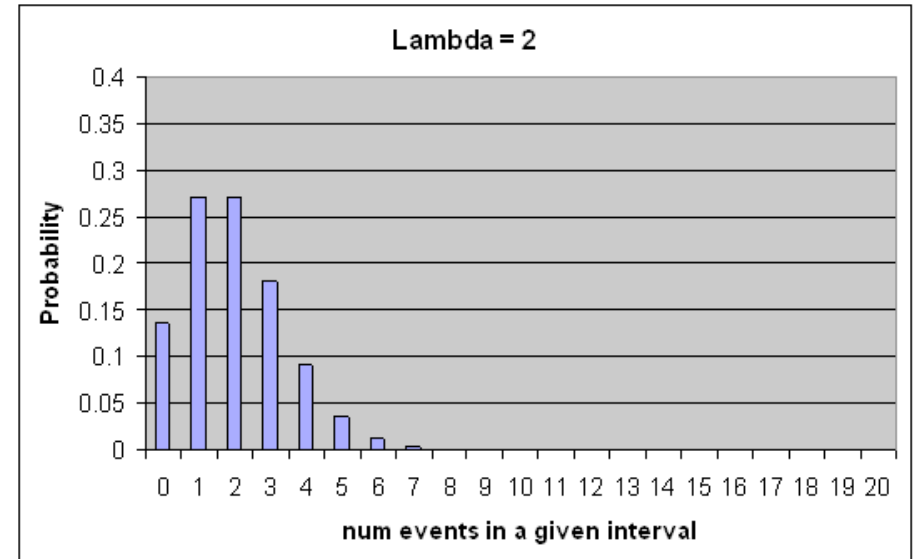
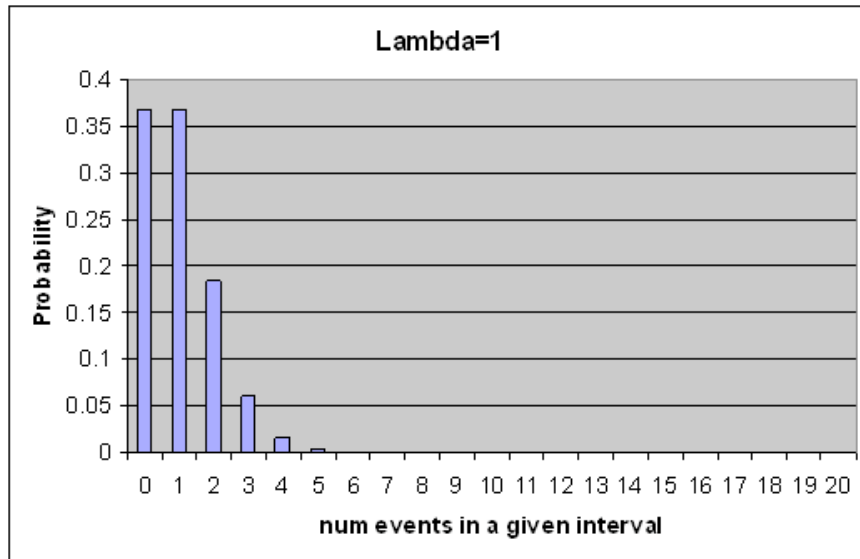
If the average number of accident occurred is 7 per week, then the probability that 2 accidents will happen during next week can be calculated as:

$$P(Y=2) = (7^2 * e^{-7}) / 2 = 0.0223$$

Some other examples of Poisson distribution are:

1. number of people going to the mall using a particular entrance in a given period of time
 2. number of typos per page in a book
- etc

The following graphs are plotted using different λ . The x-axis displays different values of y and the y-axis displays the probability of a particular y .



Lander/Waterman

Lander/Waterman suggested in their 1988 paper that the number of times a base is sequenced follows a Poisson distribution. There are two key assumptions they made:

1. reads will be randomly distributed in the genome
2. the ability to detect an overlap between two truly overlapping reads does not vary from clone to clone

The calculations detailed in the paper have been used to help planning the shotgun sequencing projects. Lets define a few variables first:

G = haploid genome length in bp

L = sequenced read length in bp

N = # reads sequenced

T = amount of overlap needed for detection in bp

Then *coverage* can be defined as the average number of times any given base in the genome is sequenced. It can be derived by dividing the total length of acquired sequences by the genome length.

In mathematical notation:

$$C = LN / G$$

Let's revisit the Poisson graphs above. Since coverage is the AVERAGE number of times a base is sequenced, it can be viewed as LAMBDA (remember, lambda is the average rate of events). Then we can interpret the X-axis as the number of times a base is sequenced. So for 1x coverage, there are many bases (about 37% of the total genome) which will not be sequenced ($x=0$). For 2x coverage, this number drops down to 14% and so on for 4x and 8x coverage. That's how researchers decide which coverage to go for.

Lets revisit the Poisson formula given above:

$$P(Y=y) = (\lambda^y * e^{-\lambda}) / y!$$

We can interpret y as *the number of times a particular base has been sequenced* and λ as *the average number of times any base in the genome is sequenced*. Since **coverage** is defined the same way as λ , we can use the value of **coverage** for λ . Given 10x coverage, if we want to calculate the probability that a base is sequenced three times, we just need to substitute 3 for y and 10 for λ into the formula:

$$\begin{aligned} P(Y=3) &= (10^3 * e^{-10}) / 3! \\ &= (1000 * 4.54 * 10^{-5}) / (3*2*1) \\ &= 0.00757 \end{aligned}$$

Note that 0.757% of bases will be sequenced only 3 times (not more than 3 or less than 3, EXACTLY 3). But often the researchers are not interested in this figure. They want to know what the probability is for a base to be sequenced 3 times or LESS. This is a more useful index in determining whether a given coverage is enough. So we want to know:

$$P(Y \leq 3)$$

Since Poisson distribution is discrete, we can break the above expression into:

$$P(Y \leq 3) = P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3)$$

Then we substitute 0, 1, 2 and 3 for y in each expression while keeping 10 for coverage level.

$$P(Y \leq 3) = (4.54 * 10^{-5}) + (4.54 * 10^{-4}) + (4.54 * 10^{-3}) / 2 + 0.00757 = 0.0103 = 1.03\%$$

About 1% of the genome is sequenced 3 times or less.

Another example of how to use the above formula is to calculate the probability that any base is NOT sequenced. We can derive the formula from the standard Poisson by substituting 0 for y .

$$\begin{aligned} P(Y=0) &= (\lambda^y * e^{-\lambda}) / y! \\ &= (C^0 * e^{-C}) / 0! \\ &= (1 * e^{-C}) / 1 \\ &= e^{-C} \end{aligned}$$

Here is a table with different coverage:

c	P0=e ^{-c}	% not sequence	% sequenced (1- P0)
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97%
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

We can derive the total gap length from the above equation:

$$\text{Total gap length} = (\% \text{ genome not sequenced}) * (\text{total genome in bp}) = e^{-c} * G$$

c	<u>target size</u>					
	50kb Ge ^{-c}	150kb Ge ^{-c}	300kb Ge ^{-c}	2Mb Ge ^{-c}	4Mb Ge ^{-c}	500Mb Ge ^{-c}
1	18,500	55,500	111,000	740,000	1,480,000	185,000,000
2	6,750	20,250	40,500	270,000	540,000	67,500,000
3	2,500	7,500	15,000	100,000	200,000	25,000,000
4	900	2,700	5,400	36,000	72,000	9,000,000
5	335	1,005	2,010	13,400	26,800	3,350,000
6	125	375	750	5,000	10,000	1,250,000
7	45	135	270	1,800	3,600	450,000
8	15	45	90	600	1,200	150,000
9	5	15	30	200	400	50,000
10	2	6	12	90	180	20,000

Number of gaps expected can also be derived:

$$\text{Number of gaps} = (\% \text{ genome not sequenced}) * (\# \text{ of reads sequenced}) = N * e^{-c}$$

150kb Target Clone:

c	Read Length					
	500			600		
	N	e ^{-c}	#Gaps=Ne ^{-c}	N	e ^{-c}	#Gaps=Ne ^{-c}
1	300	0.37	111	250	0.37	93
2	600	0.135	81	500	0.135	68
3	900	0.05	45	750	0.05	38
4	1200	0.018	22	1000	0.018	18
5	1500	0.0067	10	1250	0.0067	8
6	1800	0.0025	5	1500	0.0025	4
7	2100	0.0009	2	1750	0.0009	2
8	2400	0.0003	1	2000	0.0003	1
9	2700	0.0001	0	2250	0.0001	0
10	3000	0.000045	0	2500	0.000045	0

Sequencing progress for a 500Mb Genome

G = 500,000,000

average read length of 500 bases:

fold		Total gap length	Number of		contig	%
c	e^{-c}	in bases = Ge^{-c}	gaps = Ne^{-c}	bases/gap	length	complete
1	0.37	185,000,000	370000	500	851	63
2	0.135	67,500,000	270000	250	1620	87.5
3	0.05	25,000,000	150000	167	3167	95
4	0.018	9,000,000	72000	125	6819	98.2
5	0.0067	3,350,000	33500	100	14825	99.4
6	0.0025	1,250,000	15000	83	33250	99.75
7	0.0009	450,000	6250	72	79928	99.91
8	0.0003	150,000	2375	63	210463	99.97
9	0.0001	50,000	875	57	571371	99.99
10	0.000045	20,000	500	40	999960	99.995

Christina Chen and Jay Gertz, Department of Genetics, Washington University in St. Louis
Jan. 2005