A genetic recombination model
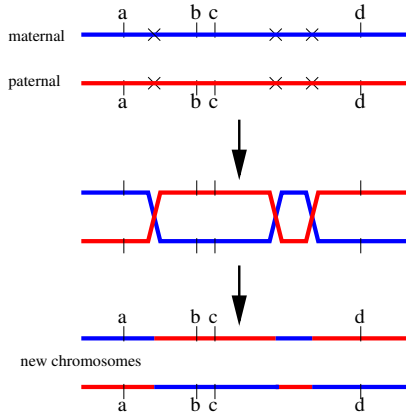
▶ In most of our cells, we have two copies of each chromosome, one inherited from our mother and one from our father. Gametes – sperm and ova – contain only one copy of each chromosome.

▶ During meiosis, the chromosomes are broken at certain random "recombination" points, to form new chromosomes out of pieces of the maternal and paternal chromosomes.

- Genes occur at particular positions along the chromosome. In early genetic research, biologists investigated the position of genes on chromosomes by studying how often they were inherited together.
- Genes on different chromosomes should be passed on independently. Genes that are close together on the same chromosome should often be passed on together. As the distance increases, the chance of being passed on together should decrease (tending to $1/2$).
- Here, genes b, c and d stay together but a is separated from them.
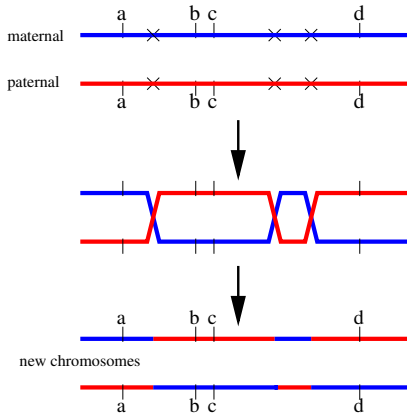
As a simple model, represent the chromosome as a continuous line, and model the recombination points as a Poisson process along it, of rate λ. Let two points a and b on the interval representing the location of two genes. Let x be the distance between them.

The probability of seeing no crossover between a and b is

$$\mathbb{P}(N(a,b] = 0) = e^{-\lambda x}.$$

The probability to see an *even number* of crossovers between a and b:

$$\sum_{k=0}^{\infty} e^{-\lambda x} \frac{(\lambda x)^{2k}}{(2k)!} = e^{-\lambda x} \left(1 + \frac{(\lambda x)^2}{2!} + \frac{(\lambda x)^4}{4!} + \cdots \right)$$

$$= e^{-\lambda x} \left(\frac{e^{\lambda x} + e^{-\lambda x}}{2} \right)$$

$$= \frac{1 + e^{-2\lambda x}}{2}.$$

If we observe that a and b are inherited together with probability $p > 1/2$, we can invert this to estimate x by

$$x = -\frac{1}{2\lambda} \log(2p - 1).$$
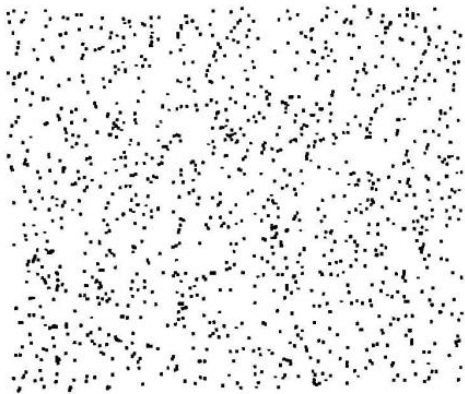
Poisson processes in higher dimensions

Consider a "process" of points in $\mathbb{R}^d$. For a set $S \subset \mathbb{R}^d$, let $N(S)$ be the number of points falling in S. The process is a Poisson process of rate λ if:

- If $S_1, S_2, \ldots, S_k$ are disjoint sets, then the counts $N(S_1), N(S_2), \ldots, N(S_k)$ are independent.
- For any set S of area a, $N(S) \sim \text{Poisson}(\lambda a)$.

Poisson processes in higher dimensions

Consider a "process" of points in $\mathbb{R}^d$. For a set $S \subset \mathbb{R}^d$, let $N(S)$ be the number of points falling in S. The process is a Poisson process of rate λ if:

- If $S_1, S_2, \ldots, S_k$ are disjoint sets, then the counts $N(S_1), N(S_2), \ldots, N(S_k)$ are independent.
- For any set S of area a, $N(S) \sim \mathsf{Poisson}(\lambda a)$.

Beginning in June, 1944, London was the target of attacks by V-1 "flying bombs" launched primarily from France. How precisely were these missiles targeted? Did the impact sites in central London cluster more than would have been expected by chance?

R. D. Clarke analysed the sites within a 12km by 12km heavily bombed region of central London. He divided the region up into 576 squares of side length 0.5km. The total number of impacts was 537.

Clarke compared the set of counts from each square with a Poisson distribution of mean $537/576$.

Distribution of bomb strikes on central London, 1944

Distribution of bomb strikes in an area of central London, 1944

Beginning in June, 1944, London was the target of attacks by V-1 "flying bombs" launched primarily from France. How precisely were these missiles targeted? Did the impact sites in central London cluster more than would have been expected by chance?

R. D. Clarke analysed the sites within a 12km by 12km heavily bombed region of central London. He divided the region up into 576 squares of side length 0.5km. The total number of impacts was 537.

Clarke compared the set of counts from each square with a Poisson distribution of mean $537/576$.

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1
	576·00	576

Review of the course

- ▶ Convergence of random variables and limit laws
  - ▶ convergence in distribution (e.g. central limit theorem)
  - ▶ convergence in probability (e.g. weak law of large numbers)
  - ▶ almost sure convergence (e.g. strong law of large numbers)
- ▶ Generating functions
  - ▶ Moment generating functions, characteristic functions. Uniqueness and continuity theorems (leading to proof of the central limit theorem).
- ▶ Joint distributions, conditional distributions, change of variables,...

- ▶ **Markov chains**
  - ▶ n-step transitions
  - ▶ communicating classes
  - ▶ periodicity
  - ▶ irreducibility
  - ▶ recurrence/transience
  - ▶ random walks on $\mathbb{Z}^d$, $d = 1, 2, 3$.
  - ▶ hitting probabilities / mean hitting times
  - ▶ mean return time: null recurrence / positive recurrence
  - ▶ stationary distributions
  - ▶ convergence to equilibrium, ergodic theorem
- ▶ **Poisson processes**
  - ▶ Independent increments property
  - ▶ Exponential interarrivals
  - ▶ Superposition and thinning

What comes next?!

Part A:
- ▶ Integration
- ▶ Statistics
- ▶ Graph Theory
- ▶ Simulation and Statistical Programming (Maths & Stats)

Part B:
- ▶ Applied Probability
- ▶ Probability, Measure and Martingales
- ▶ Continuous Martingales and Stochastic Calculus
- ▶ Mathematical Models of Financial Derivatives
- ▶ Information Theory
- ▶ Stochastic Modelling of Biological Processes
- ▶ various Statistics courses

Part C: all sorts of options in probability, statistics, genetics, stochastic analysis, combinatorics, machine learning...