Statistical Modeling and Cliodynamics

Dave Darmon and Lucia Simonelli

M3C – Tutorial 3

October 17, 2013

Overview

Why probability and statistics?

- We see the world through a glass, darkly.
  - Our models are imperfect. (Bias)
  - Our data are imperfect. (Variance)

A mundane example

- How long to walk from home to school?

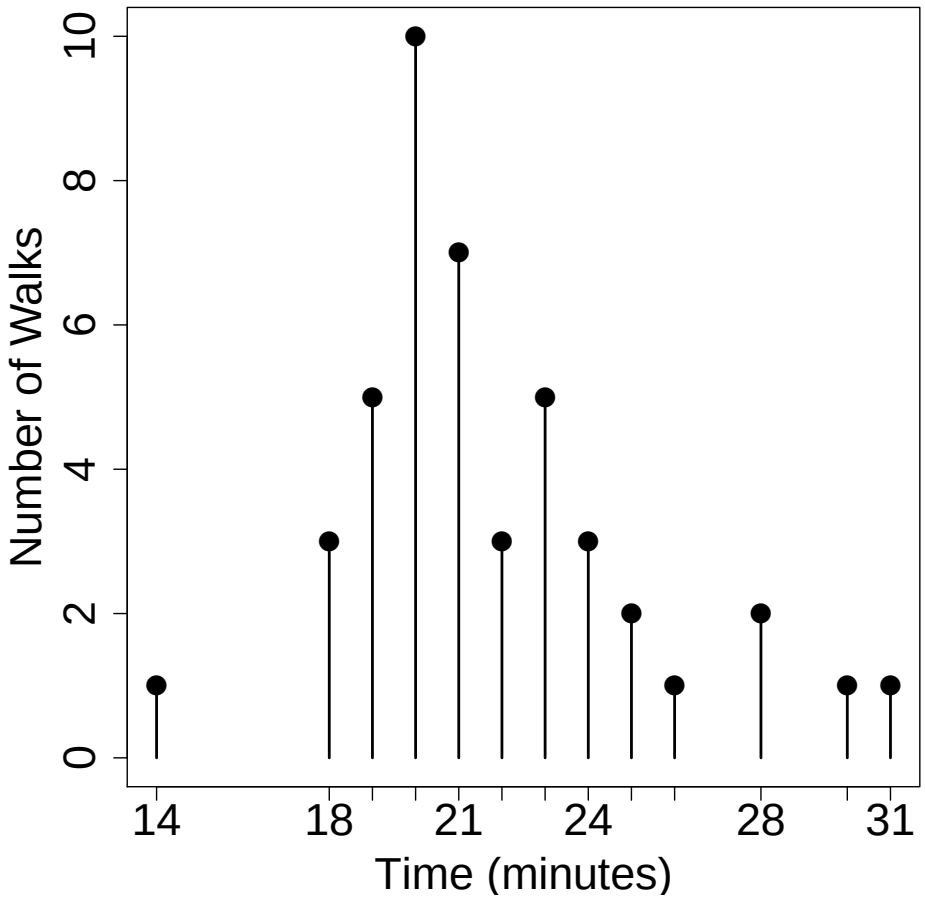


Figure: I took a walk.

A less mundane example

- Where will a hurricane hit?

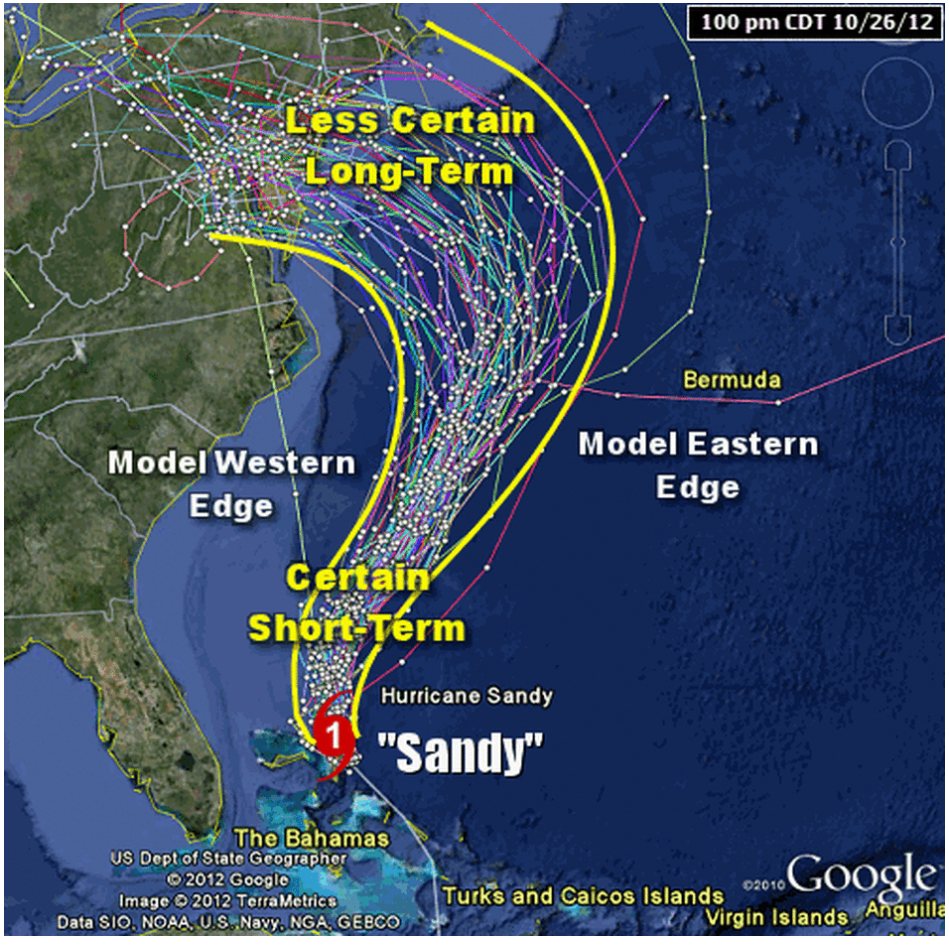


Figure: Hurricane prediction is a heady mix of modeling and data assimilation.

We're not in STAT100 anymore...

Probability Spaces

- Ω = collection of outcomes
  - examples: $\{\text{H},\text{T}\}$, $\mathbb{N}$, $[0,1]$
- A probability $\mathbb{P}$ assigns a likelihood value to each subset of Ω.

  $\mathbb{P}$ must satisfy the following properties:
  - $\mathbb{P}(\Omega) = 1$.
  - $\mathbb{P}(A) \in [0,1]$ for A a subset of Ω
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$

Random Variables

$X : \Omega \to \mathbb{R}$ (or a subset of $\mathbb{R}$)

A **random variable** is a function that can take on a set of possible different values, each with an associated probability.

- A discrete random variable takes values in a finite or countable subset of $\mathbb{R}$.
- A continuous random variable takes values on $\mathbb{R}$ or intervals of $\mathbb{R}$.

Random Variables

The **probability distribution function**, $F_x(a) = \mathbb{P}[X \leq a]$, defines the probability distribution for a given random variable X.

- For a discrete random variable
  - $F_x(a) = \sum_{x_i \leq a} \mathbb{P}[X = x_i]$
  - For discrete random variables we sometimes talk about the probability mass function $\mathbb{P}[X = a]$
  - $\sum_a \mathbb{P}[X = a] = 1$
  - Bernoulli, binomial, geometric, Poisson, etc.

- For a continuous random variable
  - $F_x(a) = \int_{-\infty}^{a} f_x(x)dx$ where $f_x(x)$ is the probability density function (pdf) satisfying:
    - $f_x(x) \geq 0$
    - $\int_{-\infty}^{\infty} f_x(x)\,dx = 1$
    - $\frac{d}{dx}F_x(x) = f_x(x)$
  - Normal, Exponential, Gamma, Beta, Cauchy, Pareto, etc.

Random Variables

Examples:

- A **Poisson** random variable, $X : \Omega \to \mathbb{N}$, with parameter λ, is a discrete random variable with probability mass function

$$\mathbb{P}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Random Variables

- A standard **Normal** (Gaussian) random variable, often denoted $N(0,1)$, is a continuous random variable with density function

$$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

and distribution

$$\mathbb{P}(X \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \, dx.$$

Random Variables

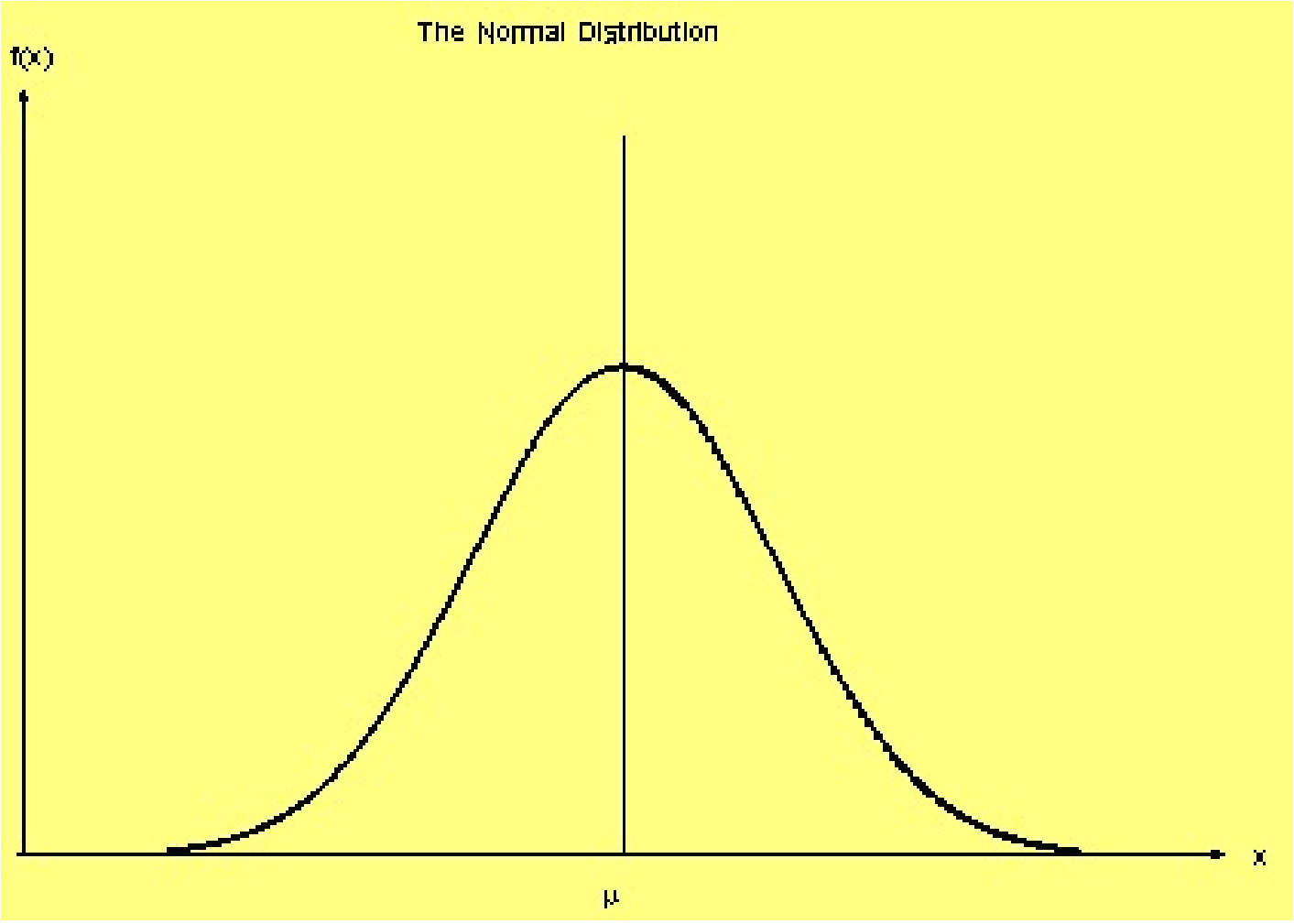


Figure: Normal Distribution

Random Variables

Expectation (Mean): weighted average of all possible values

- Discrete: $\mu = \mathbb{E}[X] = \sum_x x\mathbb{P}[X = x]$
  - If $X \sim P(\lambda)$ then $\mathbb{E}[X] = \sum_{k=0}^{\infty} k\mathbb{P}[X = k] = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda.$
- Continuous: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_x(x)\, dx$
  - If $X \sim N(0, 1)$ then $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx = 0.$

Random Variables

Variance: measure of the spread of the values of X about $\mathbb{E}[X]$.

$$\sigma^2 = Var(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- For $X \sim P(\lambda)$,

$$\text{Var(X)} = \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} - \left(\sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!}\right)^2 = \lambda$$

- For $X \sim N(0,1)$,

$$\text{Var(X)} = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx - \left(\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx\right)^2 = 1$$

Simulation Methods

These ideas can be used to simplify otherwise complicated calculations. Methods such as the **Monte Carlo Method** rely on simulation where repeated random sampling is used to obtain numerical results.

- Ex. Compute $\int_0^1 e^{-x^2}\,dx$.

Random Process

- **Random Vector**
  - multidimensional generalization of the concept of random variable
  - vector $(X_1, X_2, ..., X_k)$ of jointly distributed random variables
- **Random Process**
  - models the progression of a system over time, where the evolution is random rather than deterministic
  - a collection $(X_t : t \in T)$ where each X_t is a random variable

Random Process

Poisson Process

- A Counting Process is a random process that keeps count of the number of events that have occurred up to time t.
- A **Poisson Process**, $(N(t), t \geq 0)$, is a counting process with the following properties:
  - $N(0) = 0$
  - The process has stationary and independent increments.
  - $\mathbb{P}[N(t) = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$

What is statistics?

'Forward' Probability v. 'Inverse' Probability

- In **forward probability**, we have a probability model (distribution) and ask the consequences of that model:
  - How probable is some outcome?
  - What result do we expect, on average?
  - How variable are the outcomes?
- In **inverse probability** (statistics), we have data, and want to determine the appropriate model for that data:
  - Can we generalize from the sample to some 'population'?

What is statistics?

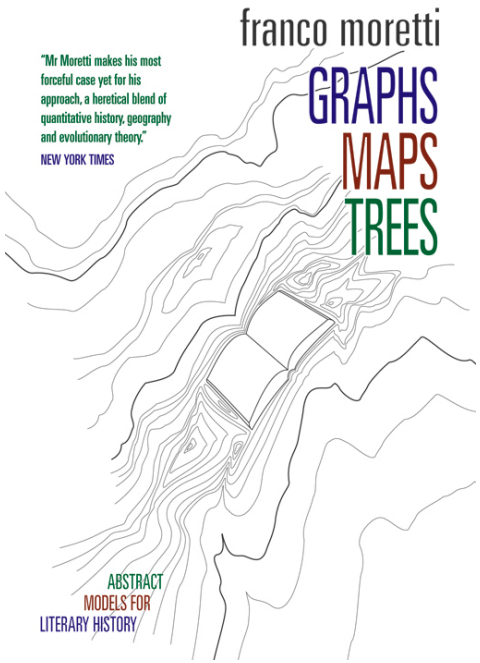
- What questions can we answer with statistical models?
  - **Parameter Estimation**: if we assume a model is correct and have data, how do we estimate the parameters?
  - **Prediction**: where is a random variable likely to end up?
  - **Hypothesis Testing / Model Checking**: which model makes the most sense to explain our data?

Cliodynamics; or, How Clumpy Are British Novels?

Cliodynamics; or, How Clumpy Are British Novels?

Figure: Franco Moretti, a literature professor at Stanford University.

Cliodynamics; or, How Clumpy Are British Novels?



- *Graphs, Maps, Trees: Abstract Models for Literary History* (2005)
- Claim from First Chapter: Genres of British novels appear together in clusters, typically separated by 25 years.

Cliodynamics; or, How Clumpy Are British Novels?

- This is a *quantitative claim*.
- We should be able to test it.
- *Cliodynamics* is a new(-ish) discipline concerned with checking such claims.
- "Clio" (Greek muse of history) + "dynamics" (study of things that vary over time)

Figure: Clio, bored in class.

Cliodynamics; or, How Clumpy Are British Novels?

The Modeling Questions: *Are you smarter than a ~~5th Grader~~ Stanford Professor?*

Given Moretti's data set of genre appearances from 1740 to 1900, devise a quantitative test of his claim that genres appear in clumps.

Cliodynamics; or, How Clumpy Are British Novels?

The data:

Genre, Begin, End
Courtship, 1740, 1820
Picareseque, 1748, 1790
Oriental, 1759, 1787
Epistolary, 1766, 1795
Sentimental, 1768, 1790
Spy, 1770, 1800

⋮

New Woman, 1888, 1899
Kailyard, 1888, 1900

Cliodynamics; or, How Clumpy Are British Novels?

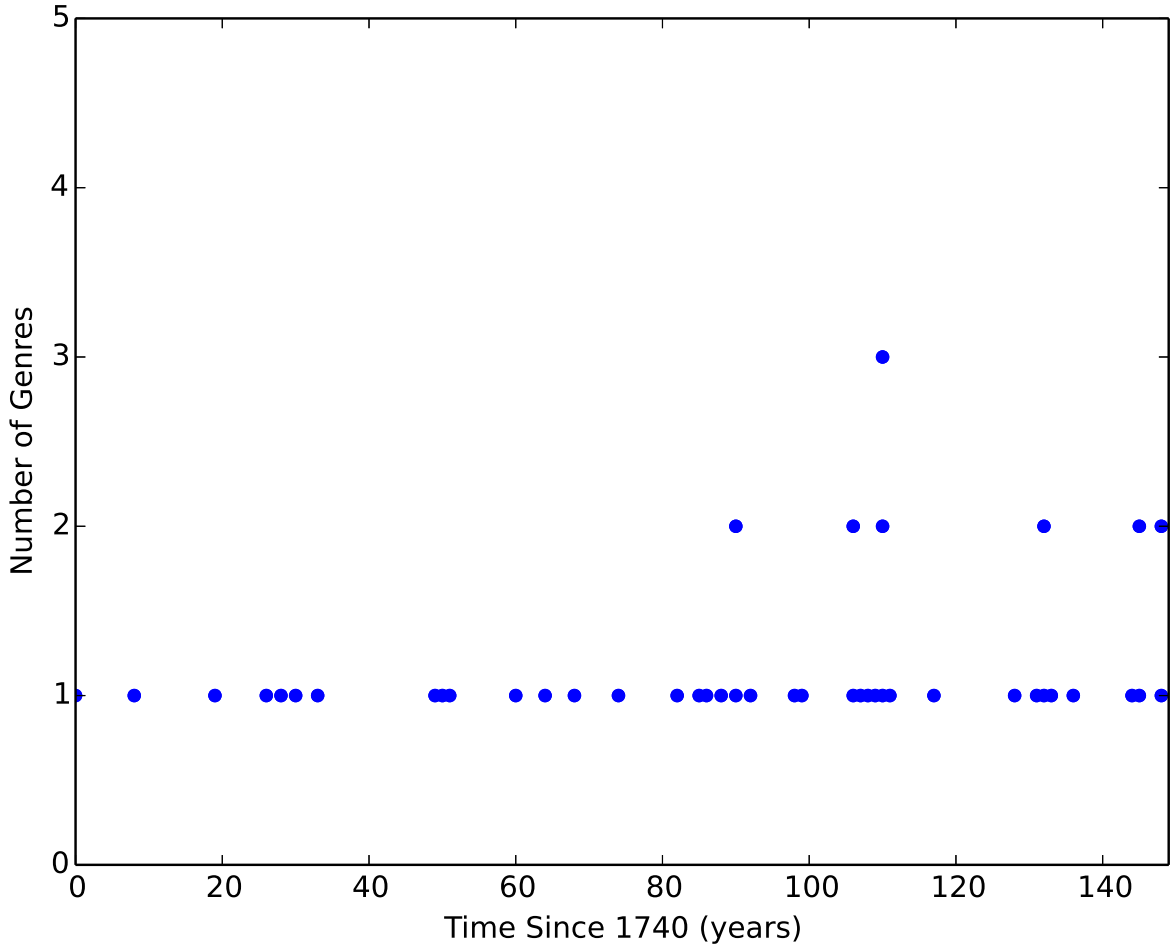


Figure: Occurrences of British novel genres over time.

Cliodynamics; or, How Clumpy Are British Novels?

If a Stanford professor says so...

- Are the genres appearing in clumps? Or are they appearing at random? It's hard to say just by looking.
- We need a way to operationalize:
  - "Random"
  - "Clumpy"

Cliodynamics; or, How Clumpy Are British Novels?

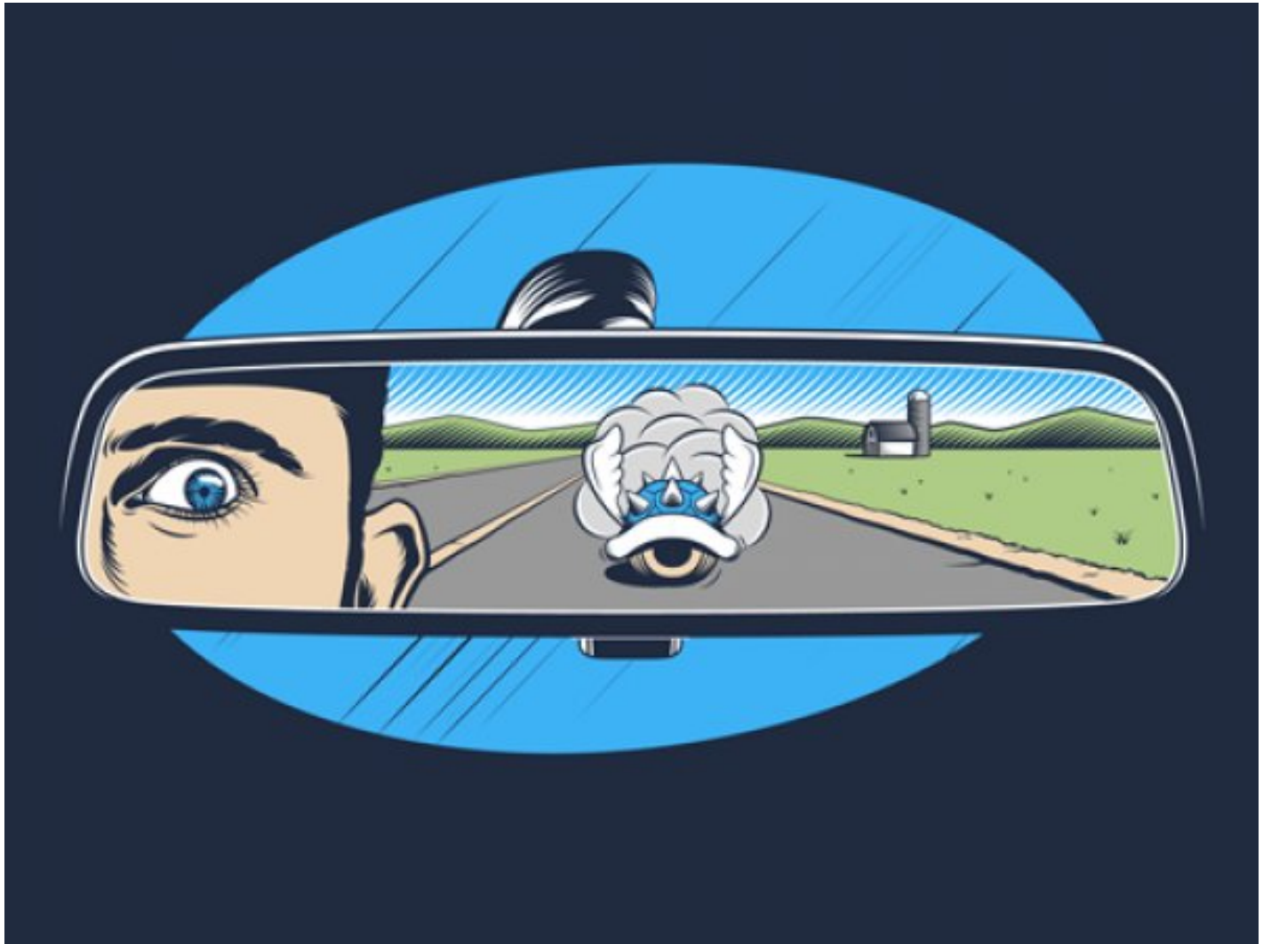


Figure: The blue shell, from Mario Kart Wii.

A Model for Randomness – The Poisson Process

- A Counting Process is a random process that keeps count of the number of events that have occurred up to time t.
- A **Poisson Process**, $(N(t), t \geq 0)$, is a counting process with the following properties:
  - $N(0) = 0$
  - The process has stationary and independent increments.
  - $\mathbb{P}[N(t) = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$

A Statistic for Clumpiness

- Let τ be the time between two successive genre appearances.
- The *coefficient of variation* of τ is defined as

$$c_v = \frac{\mathrm{Std}[\tau]}{E[\tau]}.$$

- For a Poisson process, $\tau \sim \mathrm{Exp}(\lambda)$, giving

$$c_v = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}} = 1.$$

Cliodynamics; or, How Clumpy Are British Novels?

How do the data look?

- For the novels, $\hat{c}_v = 1.06$.
- It's greater than 1 (clumpy-ish), but is it enough greater to make a difference?
- We need a way to determine if this deviation from the expected value is surprising.

Cliodynamics; or, How Clumpy Are British Novels?

The Old Way

- Lock yourself in a room with paper, a pencil, coffee, and a trash can.
- Don't leave until you've written down the sampling distribution for $\hat{c}_v$.

The New Way

"[W]e need to know what fraction of all histories of that length are at least that unlikely. I could work this out exactly, if I were willing to do some actual math, but I'm lazy, so I just had the computer simulate a million histories and evaluated all of their likelihoods."

— Cosma Shalizi

Cliodynamics; or, How Clumpy Are British Novels?

Simulation

- Sample 1 000 000 realizations from a Poisson process with rate parameter λ equal to the observed rate for the novels, $\hat{\lambda} = 0.273$.
- Compute $\hat{c}_v$ for each realization.
- Look at how $\hat{c}_v$ is distributed across the 1 000 000 realizations.
- In statistics jargon, this is called **bootstrapping** the sampling distribution of $\hat{c}_v$.

Cliodynamics; or, How Clumpy Are British Novels?

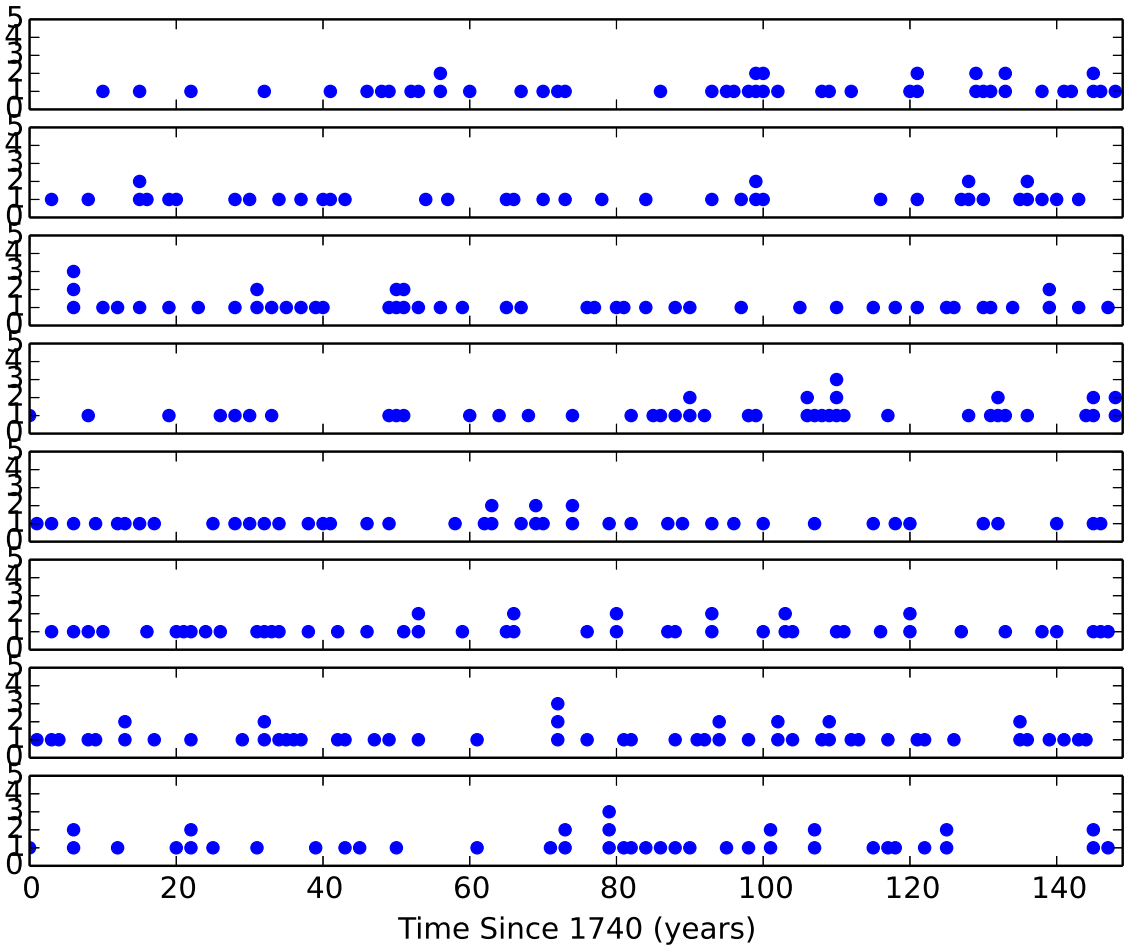


Figure: The true genre appearances and seven simulations from a Poisson process with the same rate. Can you pick out the real data?

Cliodynamics; or, How Clumpy Are British Novels?

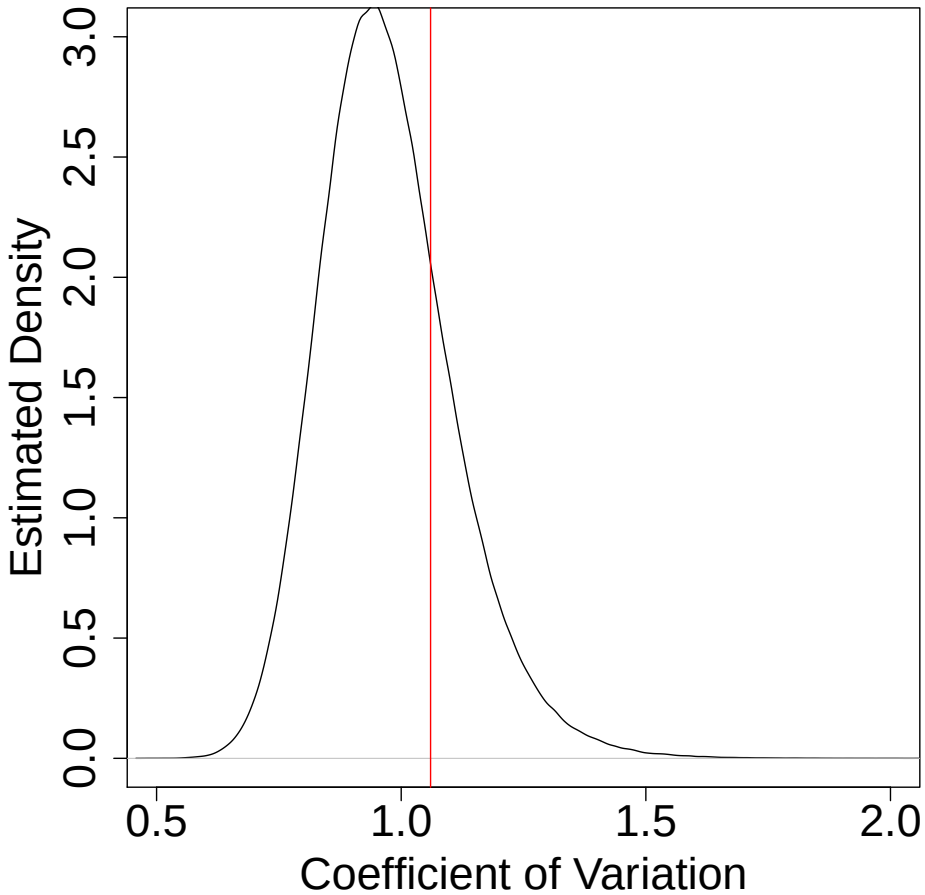


Figure: The (estimated) sampling density of the coefficient of variation, using one million realizations of a Poisson process with $\lambda = 0.273$. The red line corresponds to the observed coefficient of variation for the novels.

Cliodynamics; or, How Clumpy Are British Novels?

How random was it?

- The observed coefficient of variation was 1.06.
- Of the 1 000 000 simulations, 240 000 had coefficients of variation at least this great.
- Is something that occurs 24% of the the time for a Poisson process rare enough?

Cliodynamics; or, How Clumpy Are British Novels?

An Aside

- What *would* a 'truly clumpy' counting process look like?
- Before, we set $\lambda(t) = 0.273$ for all t. This is called a *homogeneous Poisson process*.
- Instead, set $\lambda(t) = \frac{M}{2}(\sin\left(\frac{2\pi}{T}t\right) + 1)$ where
  - $M = 2 \cdot 0.273$, the maximum rate
  - $T = 25$ years, the period.
- This is an *inhomogeneous Poisson process*. We expect the novel occurrences to 'clump' every 25 years or so.

Cliodynamics; or, How Clumpy Are British Novels?

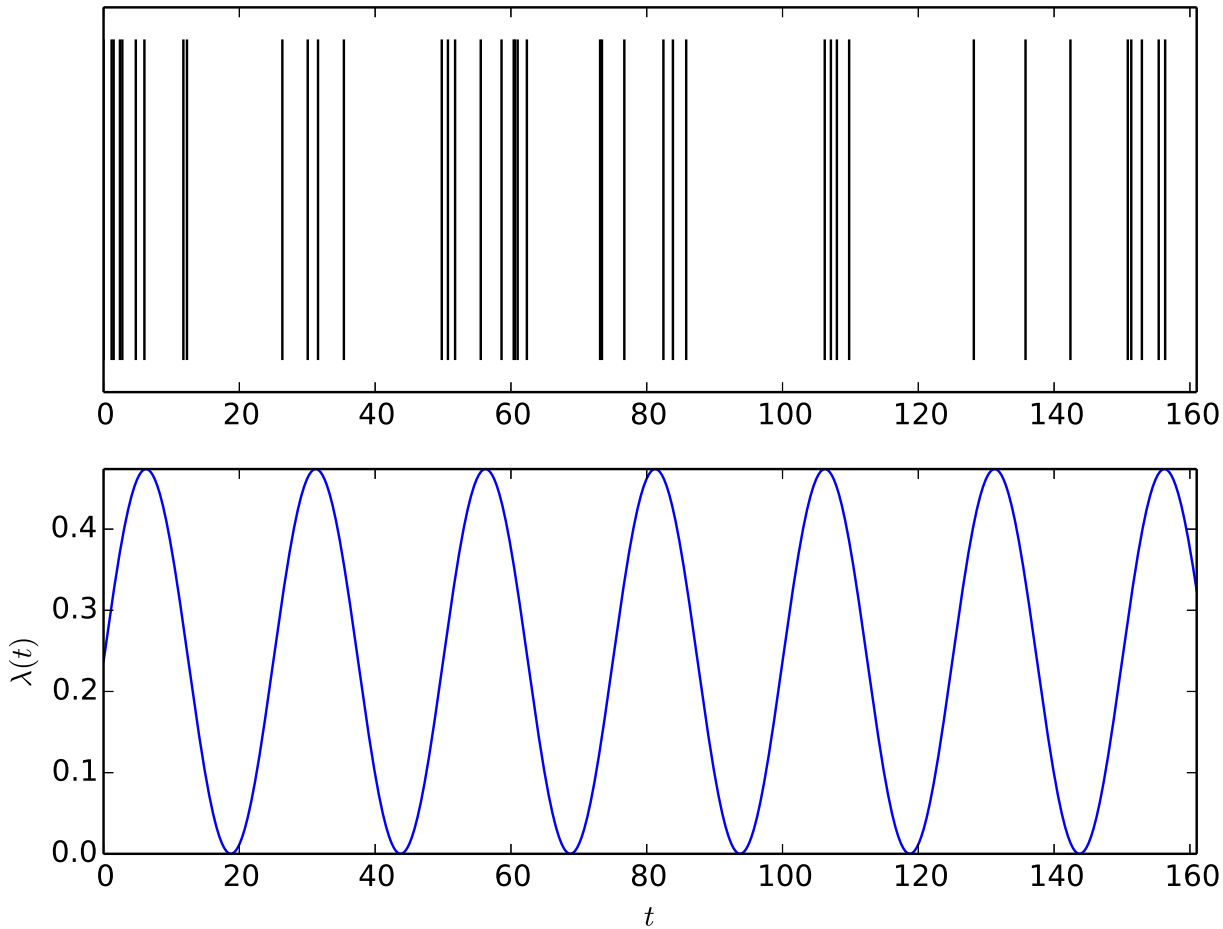


Figure: A realization from a Poisson process with $\lambda(t) = \frac{M}{2}(\sin\left(\frac{2\pi}{T}t\right) + 1)$, $M = 0.273$, $T = 25$ years.

Cliodynamics; or, How Clumpy Are British Novels?

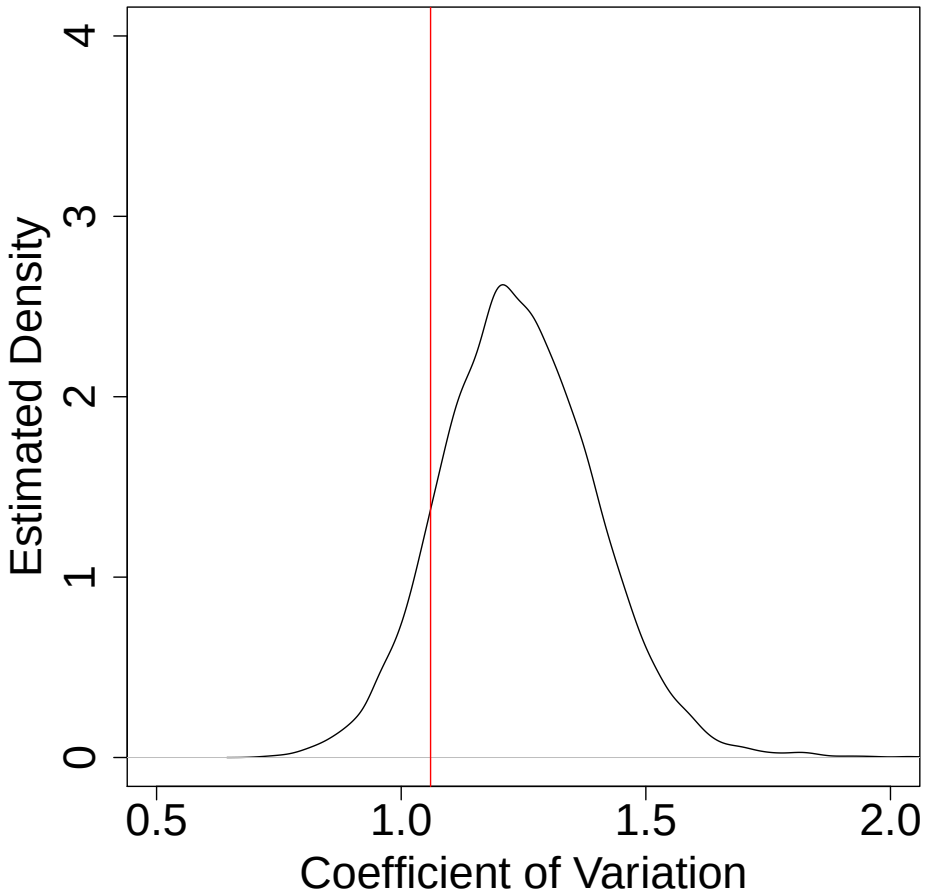


Figure: The (estimated) sampling density of $\hat{c}_v$ under the inhomogeneous Poisson process.

Why probability and statistics?

- We see the world through a glass, darkly.
  - Our models are imperfect. (Bias)
  - Our data are imperfect. (Variance)

References

Larry Wasserman, CMU

Textbook: *All of Statistics*
Blog: Normal Deviate

Cosma Shalizi, CMU

Textbook: *Advanced Data Analysis from an Elementary Point of View*
Blog: Three-Toed Sloth

Thanks!
Questions?

"This material is not necessarily sorted by topic nor graded by difficulty, although some hints, discussion and answers are given. This is because mathematics in the raw does not announce, "I am solved using such and such a technique." In most cases, half the battle is to determine how to start and which tools to apply."

— from *Mathematics by Experiment* by Borwein and Bailey