Summer School in Statistics for Astronomers & Physicists
June 5-10, 2005

Session on 'Statistical Inference for Astronomers'

Poisson Processes and Gaussian Processes

Donald Richards

Department of Statistics
Center for Astrostatistics
Penn State University

The Binomial Distribution

Bernoulli trial: A random experiment with only two possible outcomes, "success" or "failure"

$p \equiv P(\text{success})$, $q \equiv 1 - p = P(\text{failure})$

Perform n independent repetitions

X: No. of successes among all n repetitions

Possible values of X: $0, 1, 2, \ldots, n$

Probability distribution of X:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

The binomial distribution: $X \sim B(n, p)$

X: No. of micrometeorites hitting the ISS in 10,950 days

$X \sim B(n, p)$

n is large, p is small, $np = \lambda$, or $p = \frac{\lambda}{n}$

$$\begin{aligned} P(X = k) &= \frac{n!}{k!\,(n-k)}\, p^k q^{n-k} \\ &= \frac{n!}{k!\,(n-k)} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\, \frac{\lambda^k}{k!}\, \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} \end{aligned}$$

Let $n \to \infty$

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$$

$$(1 - \frac{\lambda}{n})^k \to 1, \quad (1 - \frac{\lambda}{n})^n \to e^{-\lambda}$$

Therefore

$$P(X = k) \to \frac{e^{-\lambda}\lambda^k}{k!}$$

The Poisson distribution: $Y \sim Poisson(\lambda)$ if

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The Poisson distribution is a good approximation to the binomial distribution $B(n, p)$ when n is large and p is small.

Historical note: The Poisson distribution was introduced by Siméon-Denis Poisson in his 1887 book, *Recherches sur la probabilité des jugements en matière criminelle et matière civile, précédées des règles générales du calcul des probabilités* [Researches on the Probabilities of Opinions in criminal matters and civil matters, preceded by general rules of the calculus of probabilities]

Rutherford, Chadwick, and Ellis (1920), "Radiations from radioactive substances," Cambridge, p. 172 (cited in Feller (1968), "An Introduction to Probability Theory and Its Applications," third ed., New York: Wiley.)

They counted the no. of radioactive particles emitted during 7.5-second intervals, repeating the count 2608 times. On an atomic time scale, particle emission is a rare (low probability) event.

k	Observed (N_k)	Expected
0	57	54.40
1	203	210.52
2	383	407.36
3	525	525.50
4	532	508.42
5	408	393.52
6	273	253.82
7	139	140.33
8	45	67.88
9	27	29.19
$\geq$10	16	17.08
Total	2608	2608.00

Total no. of particles: $T = \sum kN_k = 10,094$

Average no. of particles per interval:

$$T \div 2608 = 3.87$$

The right-hand column contains the values of $NP(Y = k)$ where $Y \sim Poisson(3.87)$

This is the ML method of estimating λ

The Poisson Process

Applications

A sequence of random events in time, occurring independently of each other

Emission of α-particles, micrometeorites hitting the ISS, ...

The number of earthquakes occurring during a fixed time period

The number of wars per year

The number of electrons emitted from a heated cathode during a fixed time period

Spatial distribution of stars

The distribution of galaxies (nonhomogeneous Poisson process)

Giant radio emission pulses from the Crab Nebula pulsar

Events occurring independently of each other and at random time points

N_t: No. of events observed in the interval $(0, t)$

N_t is a random variable

$N_0 = 0$ (we start observing events at $t = 0$).

Assumptions:

(a) If $t_1 < t_2 < t_3 < t_4$ then the "increments" $N_{t_2} - N_{t_1}$ and $N_{t_4} - N_{t_3}$ are independent.

(b) The distribution of $N_{t_2} - N_{t_1}$ depends only on $t_2 - t_1$, the interval length ("stationary" increments).

(c) For small h, the probability of exactly one event occurring in $(t, t+h)$ is approximately λh

(d) For small h, the probability that two or more events occur in $(t, t + h)$ is negligible

Assumption (c) means that

$$\frac{P(N_{t+h} - N_t = 1)}{h} \to \lambda \qquad \text{as } h \to 0$$

Assumption (d) means that

$$\frac{P(N_{t+h} - N_t \geq 2)}{h} \to 0 \qquad \text{as } h \to 0$$

For $h \simeq 0$, assumptions (c) and (d) imply that

$$\begin{aligned} &P(N_{t+h} - N_t = 0) \\ &\quad = 1 - P(N_{t+h} - N_t = 1) - P(N_{t+h} - N_t \geq 2) \\ &\quad \simeq 1 - \lambda h \end{aligned}$$

What is the probability distribution of N_t?

Is there a nice formula for $P(N_t = k)$, $k = 0, 1, 2, \ldots$

Remarkable result: $P(N_t = k) = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}$

Denote $P(N_t = k)$ by $R_k(t)$

$k = 0$: We shall find $P(N_t = 0)$

$N_{t+h} = 0$ means that $N_{t+h} - N_t = 0$ and $N_t = 0$

$$
\begin{aligned}
&P(N_{t+h} = 0) \\
&\quad = P(N_{t+h} - N_t = 0, N_t = 0) \\
&\quad = P(N_{t+h} - N_t = 0)\, P(N_t = 0) \text{ (indep. incrs.)} \\
&\quad \simeq P(N_t = 0)\,(1 - \lambda h)
\end{aligned}
$$

$$R_0(t + h) \simeq (1 - \lambda h)\, R_0(t)$$

$$R_0(t + h) - R_0(t) \simeq -\lambda h\, R_0(t)$$

$$\frac{R_0(t + h) - R_0(t)}{h} \simeq -\lambda\, R_0(t)$$

Take the limit as $h \to 0$: $R_0'(t) = -\lambda R_0(t)$

Solve this differential equation:

$$R_0(t) = ce^{-\lambda t}$$

Let $t \to 0$

$$R_0(t) \to R_0(0) = P(N_0 = 0) = 1$$

$$ce^{-\lambda t} \to c$$

Therefore $c = 1$

Conclusion:

$$P(N_t = 0) = R_0(t) = e^{-\lambda t}$$

$k \geq 1$: If $N_{t+h} = k$ then either

$N_t = k$, and no event occurred in $(t, t+h)$, or
$N_t = k-1$, and one event occurred in $(t, t+h)$

$$P(N_{t+h} = k) = P(N_t = k, N_{t+h} - N_t = 0) + P(N_t = k-1, N_{t+h} - N_t = 1)$$

$$P(N_t = k, N_{t+h} - N_t = 0) = P(N_t = k)P(N_{t+h} - N_t = 0) \simeq P(N_t = k)(1 - \lambda h)$$

$$P(N_t = k-1, N_{t+h} - N_t = 1) = P(N_t = k-1)P(N_{t+h} - N_t = 1) \simeq P(N_t = k-1) \cdot \lambda h$$

$$R_k(t+h) \simeq (1 - \lambda h)R_k(t) + \lambda h R_{k-1}(t)$$

$$\frac{R_k(t+h) - R_k(t)}{h} \simeq -\lambda R_k(t) + \lambda R_{k-1}(t)$$

Take the limit as $h \to 0$:

$$R_k'(t) = -\lambda R_k(t) + \lambda R_{k-1}(t)$$

$k = 1$: $R_1'(t) = -\lambda R_1(t) + \lambda R_0(t)$

$$R_1'(t) = -\lambda R_1(t) + \lambda e^{-\lambda t}$$

A first-order, linear ODE!

Solution: $R_1(t) = e^{-\lambda t}(\lambda t)$

$k = 2$: $R_2'(t) = -\lambda R_2(t) + \lambda R_1(t)$

$$R_1'(t) = -\lambda R_1(t) + \lambda e^{-\lambda t}(\lambda t)$$

Another first order, linear ODE.

Solution:

$$R_2(t) = e^{-\lambda t}\frac{(\lambda t)^2}{2!}$$

And so on ...

Probability distributions of interarrival times

We observe a Poisson process, N_t

T_1: The arrival time of the first event
T_2: The time between the arrival of the first and second events
T_3: The time between the arrival of the second and third events
etc.

$T_1, T_2, \ldots$ are continuous random variables; what are their probability distributions?

$$\begin{aligned} P(T_1 \leq t) &= P(N_t > 0) \\ &= 1 - P(N_t = 0) = 1 - e^{-\lambda t} \end{aligned}$$

T_1 has an exponential distribution

This is also the case for all the T_i's

The assumptions that increments are stationary and independent imply that, at any time t, the process has no memory and restarts itself probabilistically.

Lundgren, et al. 1995, ApJ, v.453, p.433, "Giant Pulses from the Crab Pulsar: A Joint Radio and Gamma-Ray Study"

Crab Nebula emits giant bursts of radio emission (2000 times the average pulse amplitude)

N_t: The number of giant pulses in a time period of length t

Lundgren, et al. study the distribution of interarrival times between giant pulses

They conclude that N_t is well-modeled by a Poisson process

Assignment

Collect the data on interarrival times of giant pulses from the Crab

Assume that N_t is a Poisson process with constant rate λ

Estimate λ using the method of maximum likelihood (see the example on the data of Rutherford, et al.)

The ML estimator is consistent and asymptotically normal. Using the Crab Nebula data and the asymptotic normal distribution, construct a 95% confidence interval for λ

Apply the χ^2 goodness-of-fit statistic to test the hypothesis that interarrival times of giant pulses from the Crab Nebula have an exponential distribution.

Beust, et al. 1996, A&A, v.310, p.181-198, "The β Pictoris circumstellar disk. XXII. Investigating the model of multiple cometary infalls."

Variable redshifted absorption features in the spectrum of β Pic have been attributed to comet-like bodies falling toward the star.

Some observed absorption features last too long to be consistent with the arrival of only one comet-like body.

Beust, et al. develop a Poisson process model for comet arrival.

By testing the hypothesis that interarrival times of comet-like bodies have an exponential distribution, we can verify that the Poisson model is plausible

Homogeneous case: λ not dependent on time

Nonhomogeneous case: λ is a function of t

$\lambda(t)$: The *intensity function* of N_t

As before, we assume $N - 0 = 0$

N_t is a *nonhomogeneous Poisson process* if

(a) N_t has independent increments

(b) For small h, the probability of exactly one event in $(t, t + h)$ is approximately $\lambda(t)h$

(c) For small h, the probability that two or more events occur in $(t, t + h)$ is negligible

The *mean value function* of the process:

$$m(t) = \int_0^t \lambda(u)\, du$$

By now-familiar arguments,

$$P(N_t = k) = \frac{[m(t)]^k}{k!}\, e^{-m(t)}$$

Also

$$P(N_{t+s} - N_t = k) = \frac{[m(t+s) - m(t)]^k}{k!}\, e^{-[m(t+s)-m(t)]}$$

The nonhomogeneous Poisson model is suitable for situations in which increments are independent but are not stationary.

What can we say about the interarrival times of a nonhomogeneous Poisson process? For example,

$$\begin{aligned} P(T_1 \leq t) &= P(N_t > 0) \\ &= 1 - P(N_t = 0) = 1 - e^{-m(t)} \end{aligned}$$

Therefore T_1 has an exponential distribution only if $m(t) = ct$, where c is a constant

Because the increments are independent, the T_i's are also independent

Because the increments are not stationary, the T_i's are also not stationary

Wheatland 2000, ApJ, v.536, L109-L112, "The origin of the solar flare waiting-time distribution"

The waiting-time distribution of flares is consistent with a time-dependent Poisson process

Waiting times: $S_1 = T_1$, $S_2 = T_1 + T_2$, ...

Assignment

Using the data in this article, reconstruct the observed interarrival time data from the waiting-time data

Apply to the data $t_1, t_2, t_3, \ldots$ a statistical test for independence. We can expect that we will fail to reject the null hypothesis of independence.

Plot the observed interarrival times against their labels $1, 2, 3, \ldots$, i.e., plot the points $(1, t_1)$, $(2, t_2)$, ... We should expect to see no patterns in this scatterplot.

Two-dimensional Poisson processes

Spatial Poisson processes

Boese & Doebereiner 2001, A&A v.370, 649, "Maximum likelihood estimation of single X-ray point-source parameters in ROSAT data, Astronomy and Astrophysics, ROSAT data"

Counting X-ray photons in space

Basic assumption: X-ray photons are observed one at a time

This is reasonable on an atomic time scale

Scalo & Wheeler 2002, ApJ, v.566, 723, "Astrophysical and astrobiological implications of gamma-ray burst properties"

Modeling the GRB events in the Milky Way as a spatial Poisson process

$\mathcal{P}$: Photon space

Q: A set in photon space, $Q \subseteq \mathcal{P}$

We monitor Q for photons

$N(Q)$: No. of photons observed in Q

$N(Q)$ is a random variable based on the random experiment of photon counting

$N(Q)$ is called a *spatial Poisson process* if

a. $N(Q)$ has independent increments: Whenever $Q_1, Q_2, \ldots, Q_r$ are non-overlapping subsets of photon space then $N(Q_1), N(Q_2), \ldots, N(Q_r)$ are mutually independent.

b. For each Q, $N(Q)$ has a Poisson distribution with rate λ, where λ may depend on Q.

Example: $\lambda(Q)$ is the volume of Q

The binomial and Poisson random variables: Discrete random variables

Poisson process: A counting process

We need to consider continuous processes

Gaussian Processes

X_t: A time series, indexed by time

X_t: The measurement at time t of the light curve of Algol, Rs CVn, etc.

X_t is called a *Gaussian process* if for any collection of integers $i_1, \ldots, i_n$ and real numbers $a_1, \ldots, a_n$, the random variable

$$a_1 X_{i_1} + a_2 X_{i_2} + \cdots + a_n X_{i_n}$$

has a normal distribution.

The criteria which a time series must satisfy in order to be a Gaussian process are severe

I found no papers *via* the ADS which report that an observed time series plausibly is Gaussian.

A very interesting paper is:

Leighly, "A comprehensive spectral and variability study of narrow-line Seyfert 1 galaxies observed by ASCA. I. Observations and time series analysis," ApJS, v. 125, 297-316.

Section 3.3 of this paper is highly instructive in its use of times series concepts as applied to the light curves of NLS1s, in methods for detecting stationarity, and for testing for non-Gaussianity *vs.* nonlinearity.

Leighly (Section 3.2.2) considers light curves of some NLS1s which exhibit large amplitude flares. She carries out a careful and detailed analysis of evidence that these light curves are non-Gaussian rather than nonlinear. Section 3.2.2 of her paper would serve as a superb tutorial paper on the difference between linear and Gaussian time series, and the statistical analysis of time series data which are believed to be of one type or the other.