Mathematical Biostatistics Boot Camp 2: Lecture

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

October 3, 2013

Table of contents

1. The Poisson distribution
2. Poisson approximation to the binomial
3. Person-time analysis
4. Exact tests
5. Time-to-event modeling

Pump failure data

Pump	1	2	3	4	5
Failures	5	1	5	14	3
Time	94.32	15.72	62.88	125.76	5.24

Pump	6	7	8	9	10
Failures	19	1	1	4	22
Time	31.44	1.05	1.05	2.10	10.48

From Casella and Robert, Monte Carlo Statistical Methods; first edition

The Poisson distribution

- Used to model counts
- The Poisson mass function is

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

  for $x = 0, 1, \ldots$
- The mean of this distribution is λ
- The variance of this distribution is λ
- Notice that x ranges from 0 to ∞

Some uses for the Poisson distribution

- Modeling event/time data
- Modeling radioactive decay
- Modeling survival data
- Modeling unbounded count data
- Modeling contingency tables
- Approximating binomials when n is large and p is small

Definition

- λ is the mean number of events per unit time
- Let h be very small
- Suppose we assume that
  - Prob. of an event in an interval of length h is λh while the prob. of more than one event is negligible
  - Whether or not an event occurs in one small interval does not impact whether or not an event occurs in another small interval

then, the number of events per unit time is Poisson with mean λ

Poisson approximation to the binomial

- When n is large and p is small the Poisson distribution is an accurate approximation to the binomial distribution
- Notation
    - $\lambda = np$
    - $X \sim \text{Binomial}(n, p)$, $\lambda = np$ and
    - n gets large
    - p gets small
    - λ stays constant

Proof

Rice Mathematical Statistics and Data Analysis page 41

$$
\begin{aligned}
P(X = k) &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n!}{(n-k)!n^k} \times \left(1 - \frac{\lambda}{n}\right)^{-k} \times \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \\
&\rightarrow 1 \times 1 \times \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}
$$

Notes

- That $(1 - \lambda/n)^n$ converges to $e^{-\lambda}$ for large n is a very old mathematical fact
- We can show that $\frac{n!}{(n-k)!n^k}$ goes to one easily because

$$\frac{n!}{(n-k)!n^k} = 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \ldots \times \left(1 - \frac{k-1}{n}\right)$$

each term goes to 1

Examples

Some example uses of the Poisson distribution

- deaths per day in a city
- homicides witnessed in a year
- teen pregnancies per month
- Medicare claims per day
- cases of a disease per year
- cars passing an intersection in a day
- telephone calls received by a switchboard in an hour

Some results

- If $X \sim \text{Poisson}(t\lambda)$ then
$$\frac{X - t\lambda}{\sqrt{t\lambda}} = \frac{X - \text{Mean}}{\text{SD}}$$
converges to a standard normal as $t\lambda \to \infty$
- Hence
$$\frac{(X - t\lambda)^2}{t\lambda} = \frac{(O - E)^2}{E}$$
converges to a Chi-squared with 1 degree of freedom for large $t\lambda$
- If $X \sim \text{Poisson}(t\lambda)$ then X/t is the ML estimate of λ

Proof

- Likelihood is
$$\frac{(t\lambda)^x e^{-t\lambda}}{x!}$$
- So that the log-likelihood is
$$x\log(\lambda) - t\lambda + \text{constants in } \lambda$$
- The derivative of the log likelihood is
$$x/\lambda - t$$
- Setting equal to 0 we get that $\hat\lambda = x/t$

Pump failure data

- Failures for Pump 1: 5, monitoring time: 94.32 days
- Estimate of λ, the mean number of failures per day $= 5/94.32 = .053$
- Test the hypothesis that the mean number of failures per day is larger than the industry standard, .15 events per day: $H_0 : \lambda = .15$ versus $H_a : \lambda > .15$
- $TS = (5 - 94.32 \times .15)/\sqrt{94.2 \times .15} = -2.433$
- Hence P-value is very large (.99)
- HW: Obtain a confidence interval for λ

Pump failure data

- Exact P-value can be obtained by using the Poisson distribution directly
- $P(X \geq 5)$ where $X \sim \text{Poisson}(.15 \times 94.32)$

  `ppois(5, .15 * 94.32, lower.tail = FALSE) = .995`

  very little evidence to suggest that this pump is malfunctioning
- To obtain a P-value for a two-sided alternative, double the smaller of the two one sided P-values

Pump	1	2	3	4	5
Failures	5	1	5	14	3
Time	94.32	15.72	62.88	125.76	5.24
$\hat{\lambda}$	.053	.064	.080	.111	.573
P-value	.995	.999	.908	.843	.009

Pump	6	7	8	9	10
Failures	19	1	1	4	22
Time	31.44	1.05	1.05	2.10	10.48
$\hat{\lambda}$	.604	.952	.952	1.904	2.099
P-value	1e-7	.011	.011	1e-5	2e-19

Common failure rate

- If X_i for $i = 1, \ldots, n$ are Poisson$(t_i\lambda)$ then
$$\sum X_i \sim \text{Poisson}\left(\lambda \sum t_i\right)$$
If you are willing to assume the common λ, then $\sum X_i$ contains all of the relevant information
- Clearly a common λ across pumps is not warranted. However, for illustration, assume that this is the case
- Then the total number of failures was 75
- The total monitoring time was 305.4
- The estimate of the common λ would be $75/305.4 = .246$
- Later we will discuss a test for common failure rate

Person-time analysis

OC	# of cases	# of person-years
Current users	9	2,935
Never users	239	135,130

From: Rosner Fundamentals of Biostatistics, sixth edition, page 744

- One of the most common uses of the Poisson distribution in epi is to model rates
- Estimated incidence rate amongst current users is $9/2,935 = .0038$ events per person year
- Estimated incidence rate amongst never users is $239/135,130 = .0018$ events per person year

Poisson model

- Rates are often *modeled* as Poisson
- Notice that the total number of subjects is discarded, whether the 2,935 years was comprised of 1,000 or 500 people does not come into play
- Most useful for rare events (though we'll discuss another motivation for the Poisson model later)
  - $\lambda = .3$ events per year
  - Followed 10 people for a total of 40 person-years
  - Expected number of deaths is $.3 \times 40 = 12$, larger than our sample
- λ is assumed constant over time

Comparing two Poisson means

- Want to test $H_0 : \lambda_1 = \lambda_2 = \lambda$
- Observed counts x_1, x_2 Person-times t_1, t_2
- Estimate of λ under the null hypothesis is $\hat\lambda = \frac{x_1 + x_2}{t_1 + t_2}$
- Estimated expected count in Group 1 under H_0

$$E_1 = \hat\lambda t_1 = (x_1 + x_2)\frac{t_1}{t_1 + t_2}$$

- Estimated expected count in Group 2 under H_0

$$E_2 = \hat\lambda t_2 = (x_1 + x_2)\frac{t_2}{t_1 + t_2}$$

Notes

- Test statistic

$$TS = \sum \frac{(O - E)^2}{E} = \frac{(x_1 - E_1)^2}{E_1} + \frac{(x_2 - E_2)^2}{E_2}$$

  follows a Chi-squared distribution with 1 df

- Equivalent computational form

$$TS = \frac{(X_1 - E_1)^2}{V_1}$$

  where

$$V_1 = (x_1 + x_2) t_1 t_2 / (t_1 + t_2)^2$$

OC example

- $x_1 = 9, \; t_1 = 2,935$
- $x_2 = 239, \; t_2 = 135,130$
- $E_1 = (9 + 239) \times \frac{2{,}935}{2{,}935 + 135{,}130} = 5.27$
- $V_1 = (9 + 239) \times \frac{2{,}935 \times 135{,}130}{(2{,}935 + 135{,}130)^2} = 5.16$

$$TS = \frac{(9 - 5.27)^2}{5.16} = 2.70$$

$\text{P-value} = .100$

Estimating the relative rate

- Relative rate λ_1/λ_2
- Estimate $(x_1/t_1)/(x_2/t_2)$
- Standard error for the **log** relative rate estimate

$$\sqrt{\frac{1}{x_1} + \frac{1}{x_2}}$$

- For the OC example the estimated log relative rate is .550
- The standard error is $\sqrt{\frac{1}{9} + \frac{1}{239}} = .340$
- 95% CI for the log relative rate is

$$.550 \pm 1.96 \times .340 = (-.115, 1.26)$$

Exact test

- HW: Extend the above test to more than two groups
- Suppose that the observed counts were low
- We obtain an exact test by conditioning on the sum of the counts
- That is, we make use of the fact that

$$X_1 \mid X_1 + X_2 = z \sim \text{Binomial}\left(\frac{t_1}{t_1 + t_2}, z\right)$$

Alternate motivation for Poisson model

- Another way to motivate the Poisson model is to show that it is *likelihood equivalent* to a plausible model
- We show that the Poisson model is likelihood equivalent to a model that specifies failure times as being independent exponentials
- Whenever the independent exponential model is reasonable, then so is the Poisson model, regardless of how large or small the rate or sample size is
- The likelihood equivalence implies that likelihood methods apply

Recall

- The likelihood is the density viewed as a function of the parameter
- The likelihood summarizes the evidence in the data about the parameter
- If $X \sim \mathrm{Poisson}(t\lambda)$ then the likelihood for λ is

$$\mathcal{L}(\lambda) = \frac{(t\lambda)^x e^{-t\lambda}}{x!} \propto \lambda^x e^{-t\lambda}$$

- When you have independent observations, the likelihood is the product of the likelihood for each observation

Likelihood equivalence with exponential model

- Suppose there is no censoring (every person followed until an event)
- We model each person's time until failure as independent exponentials: $Y_i \sim \mathrm{Exponential}(\lambda)$
- Each subject contributes $\lambda e^{-y_i \lambda}$ to the likelihood

$$\prod_{i=1}^{n} \lambda e^{-y_i \lambda} = \lambda^n \exp\left(-\lambda \sum y_i\right)$$

- Here n is the number of events and $\sum y_i$ is the total person-time
- Same likelihood as if we specify $n \sim \mathrm{Poisson}(\lambda \sum y_i)$!

Likelihood equivalence with censoring

- Suppose now that the study ended before events were observed on some subjects
- Likelihood contribution for each event is $\lambda e^{-y_i \lambda}$
- Likelihood contribution for each censored observation is

$$P(Y \geq y_i; \lambda) = \int_{y_i}^{\infty} \lambda e^{-u\lambda} du = e^{-y_i \lambda}$$

The contribution is this integral b/c we don't really know the event time for that person, only that it was later than y_i

- Combining the censored and non-censored observations, we have

$$\left\{\prod_{\text{non}-\text{censored } i} \lambda e^{-y_i\lambda}\right\} \times \left\{\prod_{\text{censored } i} e^{-y_i\lambda}\right\}$$

x events (and hence $n - x$ censored observations)

$$= \lambda^x \exp\left(-\lambda \sum_{i=1}^n y_i\right)$$

- x is the total number of events and $\sum_{i=1}^n y_i$ is the total amount of person-time
- This is exactly the same likelihood as modeling $X \sim \text{Poisson}(\lambda \sum_{i=1}^n y_i)$