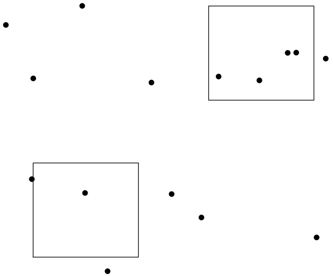
Count data and the Poisson distribution

Patrick Breheny

November 5

Poisson model: A spatial illustration

- Consider the following illustration, which might represent, for example, the position of stars in the sky, or cases of disease in a certain geographical region:



- Now suppose we specify some sets, say, A_1 and A_2, and let $N(A)$ denote the random variable that counts the number of cases in set A

The Poisson distribution

- Suppose we make two assumptions on the random process generating these points:
  - For any set A, $\mathrm{E}\{N(A)\} = \lambda\,|A|$, where $|A|$ represents the size of set A (here, its area)
  - For any sets $A_1, A_2, \ldots$ that do not overlap, $N(A_1), N(A_2), \ldots$ are independent
- **Theorem:** Under the two assumptions above, the probability mass function for $X = N(A)$ is

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

  where A is any set with $|A| = 1$
- A random variable with the above distribution is said to follow a *Poisson distribution* with rate λ: $X \sim \mathrm{Pois}(\lambda)$

Uses for the Poisson distribution

- Our illustration here was two-dimension and spatial, but the idea applies equally well to other dimensions:
    - One dimension: Modeling the number of mutations in a region of DNA
    - One dimension (time): Modeling the number of murders in London over stretches of time; modeling the numbers of hits a website receives
    - Three dimensions (volume): Modeling the number of parasites in a lake
    - Three dimensions (space-time): Modeling spatiotemporal phenomena such as the cases of a disease by location and time
- The Poisson distribution is widely used to model counts of events; in medicine and public health, the counts we are usually interested in is cases of a disease

Additivity and variance of the Poisson distribution

- From our construction of the Poisson distribution, it should be obvious that the sum of two independent Poisson random variables also follow a Poisson distribution
- Formally, if $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$ and $X \amalg Y$, then $X + Y \sim \text{Pois}(\lambda + \mu)$
- In addition, it can be shown that the variance of the Poisson distribution is equal to its mean
- Thus, if $X \sim \text{Pois}(\lambda)$, $E(X) = \lambda$ and $\text{Var}(X) = \lambda$

Marginal distributions conditional on the total

- A particularly interesting and useful feature of the Poisson distribution is that for two independent Poisson random variables, conditional on the total number of events between the two, the distribution of either count follows a Binomial distribution
- **Theorem:** Let $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$ and $X \amalg Y$. Then

$$X|X+Y=n \sim \text{Binom}\left(n, \frac{\lambda}{\lambda+\mu}\right)$$

- This fact makes it particularly easy to test hypotheses and construct confidence intervals for two-sample count data using methods we have already developed for the binomial distribution

Example

- The General Social Survey (GSS) is a sociological survey of United States residents carried out every other year by the National Opinion Research Center at the University of Chicago
- The survey collected data on 155 women who were 40 years of age or older in the 1990s
- Among the 155 women were 111 whose highest educational level was less than a bachelor's degree (these women had a total of 217 children) and 44 women with at least a bachelor's degree (these women had a total of 66 children)

Setup

- Viewing the number of children per woman as following a Poisson distribution, we have

$$\begin{aligned} &\text{Low education:} && X \sim \text{Pois}(\lambda \cdot 111) \\ &\text{High education:} && Y \sim \text{Pois}(\mu \cdot 44), \end{aligned}$$

  where λ is the expected number of children per woman in the low-education group and μ is the expected number of children per woman in the high-education group

- Thus, conditional on the fact that we observed a total of $n = 283$ children born, $X \sim \text{Binom}(n, \pi)$, where

$$\pi = \frac{111\lambda}{111\lambda + 44\mu}$$

Hypothesis testing

- So, suppose we wish to test the hypothesis that $\lambda = \mu$; this is equivalent to testing whether

$$\pi = \frac{111}{111 + 44} = 0.716$$

- Carrying out an exact binomial test of $H_0 : X \sim \text{Binom}(n = 283, \pi = 0.716)$ with $X = 217$ yields a p-value of 0.06; so, somewhere in the neighborhood of modestly significant evidence that these two groups of women have different numbers of children on average
- Note the conceptual similarity here between this test and Fisher's exact test, in that both approaches condition on an ancillary statistic (the total count) that contains no information about the quantity we are interested in (the difference in rates)

Confidence interval

- In terms of measuring the relationship between the two rates, a natural choice is the rate ratio λ/μ
- The observed rate ratio is $(217/111)/(66/44) = 1.3$; on average, women in the low-education group had 30% more children
- How to get a confidence interval?
- It is straightforward, of course, to obtain a confidence interval for π (using, say, the Clopper-Pearson approach); but what does this tell us about the rate ratio λ/μ?

Confidence interval (cont'd)

- Expressing the relationship between Poisson rates and the Binomial proportion in terms of the odds, we have

$$\frac{111\lambda}{44\mu} = \frac{\pi}{1-\pi},$$

and therefore

$$\frac{\lambda}{\mu} = \frac{44}{111}\frac{\pi}{1-\pi}$$

- Plugging the endpoints of the Clopper-Pearson interval for π, we obtain the interval $[0.99, 1.74]$
- Again, it should come as no surprise that this interval barely overlaps with 1; the data is consistent with no difference between the groups as well as consistent with the possibility that the low-education group has up to 70% more children

Confidence interval

- Hypothesis tests and confidence intervals for single Poisson rates are straightforward
- For testing the hypothesis $H_0 : \lambda = \lambda_0$, we can easily calculate p-values from the CDF of the Poisson distribution
- To obtain confidence intervals, we can simply invert the above test to obtain a region of plausible values for λ exactly as we did in constructing the Clopper-Pearson interval:
  - Low education: $[1.70, 2.23]$
  - High education: $[1.16, 1.91]$

Summary

- If event counts are homogeneous and independent, they will follow a Poisson distribution
- Important properties of the Poisson distribution:
  - If $X \sim \text{Pois}(\lambda)$, $E(X) = \lambda$ and $\text{Var}(X) = \lambda$
  - If $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$ and $X \amalg Y$, then $X + Y \sim \text{Pois}(\lambda + \mu)$
- Hypothesis testing and confidence intervals for comparing two Poisson rates can be accomplished using the fact that, conditional on the total count n, $X \sim \text{Binom}(n, \lambda/(\lambda + \mu))$