Poisson Random Variables
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-10-01 11:05

Poisson

- Chapter 6.5 text
- Two main applications:
  - Modeling counts of discrete events in space or time
  - Approximation to the Binomial distribution for large N and small p

Poisson - Examples

- Number of abnormal cells in a fixed area of a histological slide
- Count of bacteria surviving treatment in a fixed volume of bacterial suspension
- Number of white blood cells in a drop of blood
- Number of new breast cancer cases registered per month by the National Cancer Registry
- Number of live births in Greater London during the month of January

Poisson

- Two assumptions required for Poisson distribution to be an appropriate model:
  - The number of events occurring in one part of the continuum (space, time) should be statistically independent of the number of events occurring in another part of the continuum
  - The expected number of counts in a given part of the continuum should approach zero as its size approaches zero

Poisson

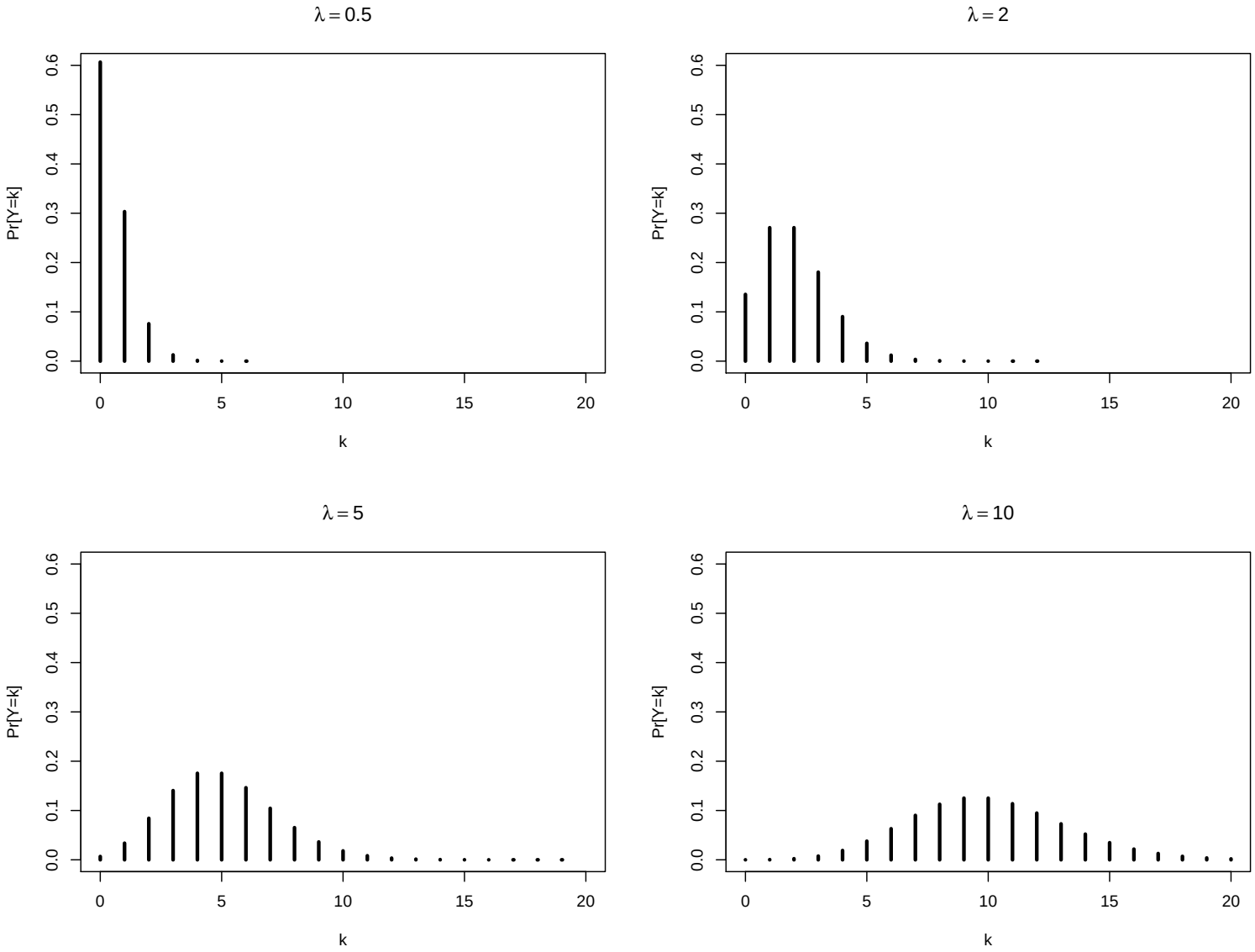
- The Poisson distribution is characterized by one parameter, λ
- $Y \sim \text{Poisson}(\lambda)$, probability mass function

$$\Pr[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

- $Y \in \{0, 1, 2, \ldots\}$
- The parameter λ is both the mean and variance

$$E(Y) = V(Y) = \lambda$$

Poisson PMF



Poisson and Binomial

- Suppose $X \sim \text{Binomial}(N, \pi)$ and $Y \sim \text{Poisson}(\lambda)$ with $\lambda = N\pi$
- Then for N large and π small

$$\Pr[X = x] \approx \Pr[Y = x]$$

i.e.

$$\binom{N}{x} \pi^x (1-\pi)^{N-x} \approx \frac{e^{-N\pi}(N\pi)^x}{x!}$$

- Rule of thumb: $\pi \leq .1$ and $N \geq 20$

Poisson and Binomial

- Table 6.6 text

	Binomial PMF				Poisson
	$N = 10$	$N = 20$	$N = 40$	$N = 1000$	PMF
k	$\pi = 0.20$	$\pi = 0.10$	$\pi = 0.05$	$\pi = 0.002$	$\lambda = 2$
0	0.1074	0.1216	0.1285	0.1351	0.1353
1	0.2684	0.2702	0.2706	0.2707	0.2707
2	0.3020	0.2852	0.2777	0.2709	0.2707
3	0.2013	0.1901	0.1851	0.1806	0.1804
4	0.0881	0.0898	0.0901	0.0902	0.0902
⋮	⋮	⋮	⋮	⋮	⋮

Poisson and Binomial

- Sketch of proof: Suppose on average μ events expected to occur over some fixed time interval
- Divide interval into N subintervals small enough such that the probability of two events occurring in the same subinterval is v unlikely
- Then the N subintervals approximate a sequence of N Bernoulli trials with success prob μ/N

Poisson and Binomial

- Thus the probability of observe exactly x events in the N subintervals is

$$\frac{N(N-1)\cdots(N-x+1)}{x!}\left(\frac{\mu}{N}\right)^x\left(1-\frac{\mu}{N}\right)^{N-x} \quad (1)$$

- As $N \to \infty$,

$$N(N-1)\cdots(N-x+1) \approx N^x$$

and

$$\left(1-\frac{\mu}{N}\right)^{N-x} \approx \left(1-\frac{\mu}{N}\right)^{N} \to e^{-\mu}$$

- Thus (1) approx equals

$$\frac{N^x}{x!}\left(\frac{\mu}{N}\right)^x e^{-\mu} = \frac{e^{-\mu}\mu^x}{x!}$$

Exact Confidence Intervals

- Cf Note 6.8 text (page 195)
- Given y occurrences, an exact $(1-\alpha)$ 100% CI for λ is

$$\left[\frac{1}{2}\chi^2_{\alpha/2;2y},\ \frac{1}{2}\chi^2_{1-\alpha/2;2(y+1)}\right]$$

Normal Approximations

- If $Y \sim$Poisson(λ) and λ large (say ≥ 100), then

$$Y \sim N(\lambda, \lambda)$$

- Thus approx $(1-\alpha)$ % CI

$$Y \pm z_{1-\alpha/2}\sqrt{Y}$$

- A better approximation arises from

$$\sqrt{Y} \sim N(\sqrt{\lambda}, \frac{1}{4})$$

- For $\lambda \geq 30$, approx CI for $\sqrt{\lambda}$

$$\sqrt{Y} \pm \frac{z_{1-\alpha/2}}{2}$$

Sum of Poisson Random Variables

- If $Y_1, Y_2, \ldots, Y_N$ iid Poisson(λ), then

$$\sum_{i=1}^{N} Y_i \sim \text{Poisson}(N\lambda)$$

- Estimator for λ

$$\hat{\lambda} = \frac{1}{N} \sum_i Y_i$$

- If (L, U) is $(1-\alpha)\%$ CI for $N\lambda$, then $(L/N, U/N)$ is $(1-\alpha)\%$ CI for λ. E.g.,

$$\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\sum_i Y_i / N^2}$$

Example 6.20

- Number of bacterial colonies per plate: 72, 69, 63, 59, 59, 53, 51
- Sum 426, mean 60.86
- Exact 95% CI for 7λ

$$\left[\frac{1}{2}\chi^2_{.025;2*426},\ \frac{1}{2}\chi^2_{.975;2*427}\right] = [386.50, 468.44]$$

Example 6.20

- Normal approximations:

$$426 \pm z_{.975} * \sqrt{426} = [385.55, 466.45]$$

$$\left[\left(\sqrt{426} - \frac{z_{.975}}{2}\right)^2, \left(\sqrt{426} + \frac{z_{.975}}{2}\right)^2\right] = [386.51, 467.41]$$

- Divide endpoints by $N = 7$ to get 95% CI for λ

Rules of Thumb

- For $\alpha = 0.05$,

$$\sqrt{Y} \pm \frac{z_{1-\alpha/2}}{2} \approx \sqrt{Y} \pm 1$$

  implying approx 95% CI

$$\left[\left(\sqrt{Y} - 1\right)^2, \left(\sqrt{Y} + 1\right)^2\right]$$

- If we observe $y = 0$, a two-sided 90% CI for λ is

$$[0, \frac{1}{2}\chi^2_{.95,2}] = [0, 3.00]$$

- Thus if we observed 0 events out of N trials, the approx upper bound on a two-sided 90% CI is $3/N$

Homogeneity Test

- Often, observed counts exhibit larger variance than expected under the Poisson model; *over-dispersion*
- This may be caused by heterogeneity in the λ's
- Want to test

$$H_0 : X_1, X_2, \ldots, X_k \sim \text{ Poisson}(\lambda)$$

Homogeneity Test

- Construct χ^2 GOF test using the following result

- Suppose $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, 2, \ldots, k$
- Then the conditional distribution of $(X_1, \ldots, X_k)$ given $\sum_i X_i = N$ is multinomial with cell probabilities

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k} \text{ for } i = 1, 2, \ldots, k$$

Homogeneity Test

- Under H_0,

$$H_0 : X_1, X_2, \ldots, X_k \sim \text{ Poisson}(\lambda)$$

the test statistic

$$T = \frac{\sum_i (X_i - \bar{X})^2}{\bar{X}} \sim \chi^2_{k-1}$$

- Equivalent form

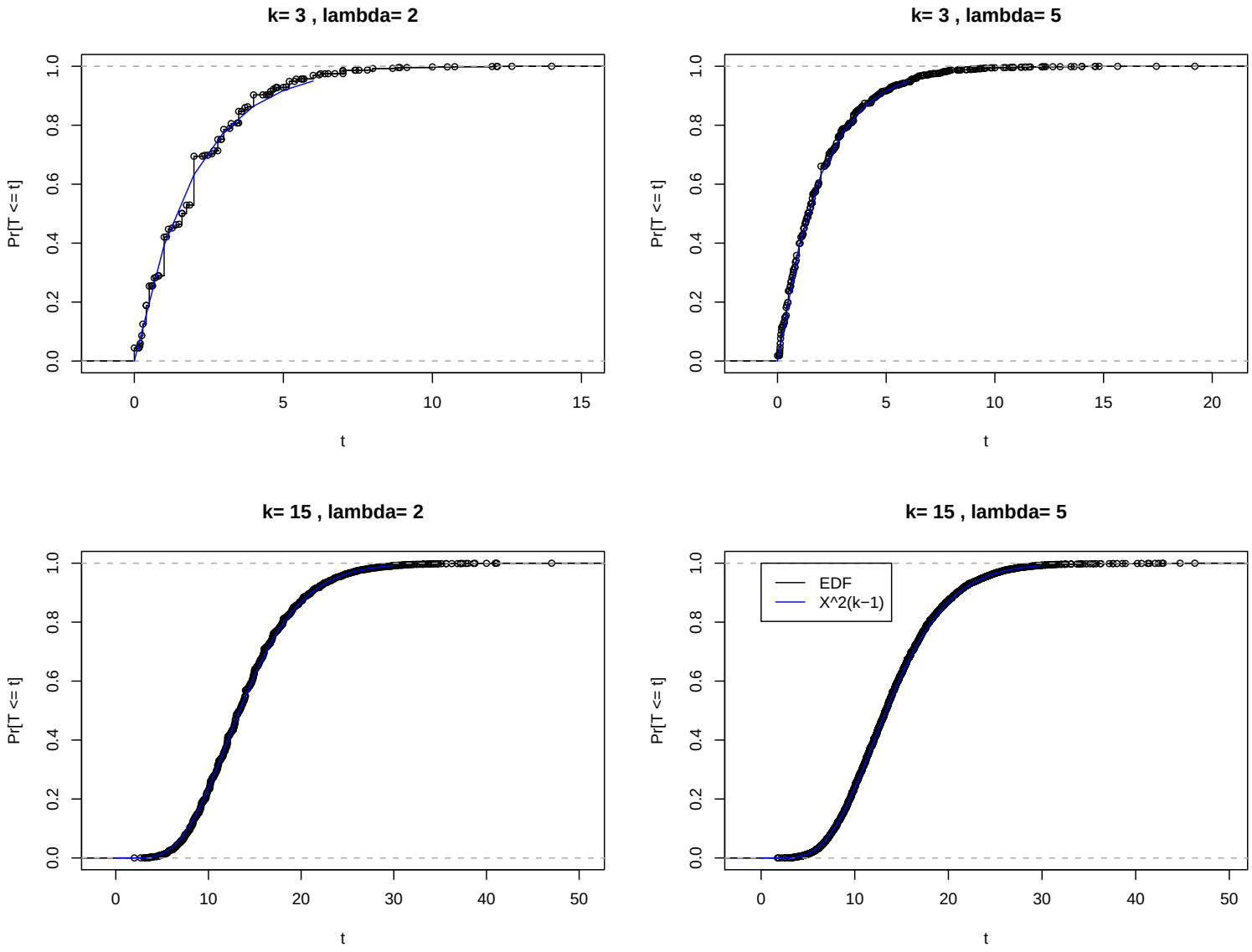
$$T = \frac{(k-1)s^2}{\bar{X}}$$

- *Poisson homogeneity/heterogeneity/dispersion test*

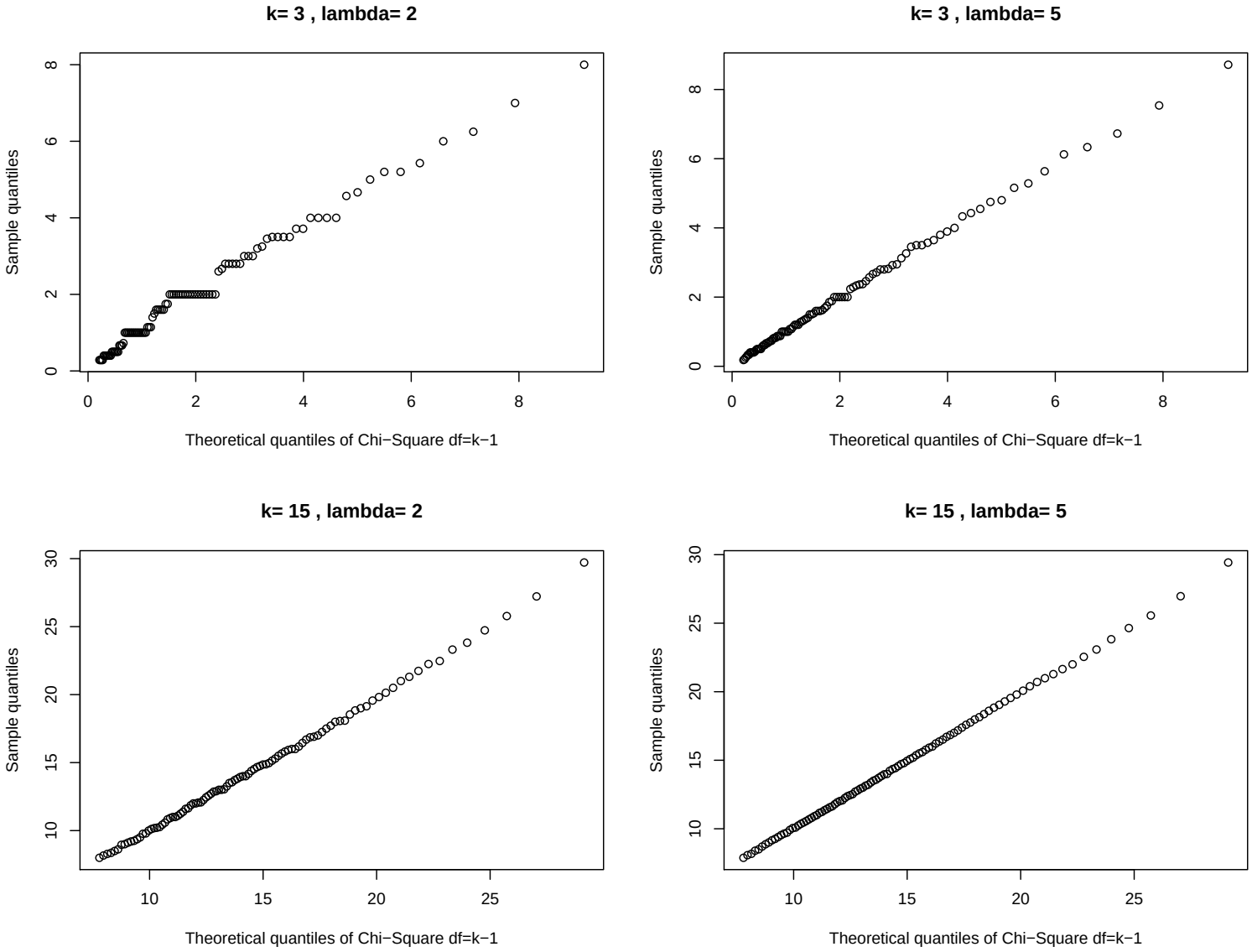
Homogeneity Test

- χ^2 approx improves as λ, k get large
- Recommendation (Armitage and Berry 1987)
  1. $\bar{X} \geq 5$, or
  2. $\bar{X} \geq 2$ and $k > 15$

Homogeneity Test: Simulation Study



Homogeneity Test: Simulation Study



Homogeneity Test

- Example 6.20

$$k = 7,\ \bar{X} = 60.86,\ s_X = 7.7552$$

  implying

$$T = \frac{6 * (7.7552)^2}{60.86} = 5.93$$

- Since $\Pr[\chi^2_6 > 5.93] = 0.43$, fail to reject H_0

Hypergeometric

- Suppose $X_{ij} \sim \text{Poisson}(\lambda_{ij})$ for $i, j = 1, 2$
- Then the distribution of X_{11} conditional on $X_{i1}+X_{i2} = m_i$ for $i = 1, 2$ and $X_{1j} + X_{2j} = n_j$ for $j = 1, 2$ is non-central hypergeometric with non-centrality parameter $\theta = \lambda_{11}\lambda_{22}/\lambda_{12}\lambda_{21}$

Negative-Binomial

- Overdispersion may be due to heterogeneity of λ's
- I.e, λ is no longer a constant, but a random variable
- If λ follows a gamma distribution, then the counts follow a negative binomial distribution
- NB allows for variance to be proportional to the mean