BIOS 662 Fall 2018

Poisson Random Variables

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

The Poisson Distribution

- Chapter 6.5 of the 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

The Poisson Distribution

- Two assumptions required for the 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

The Poisson Distribution

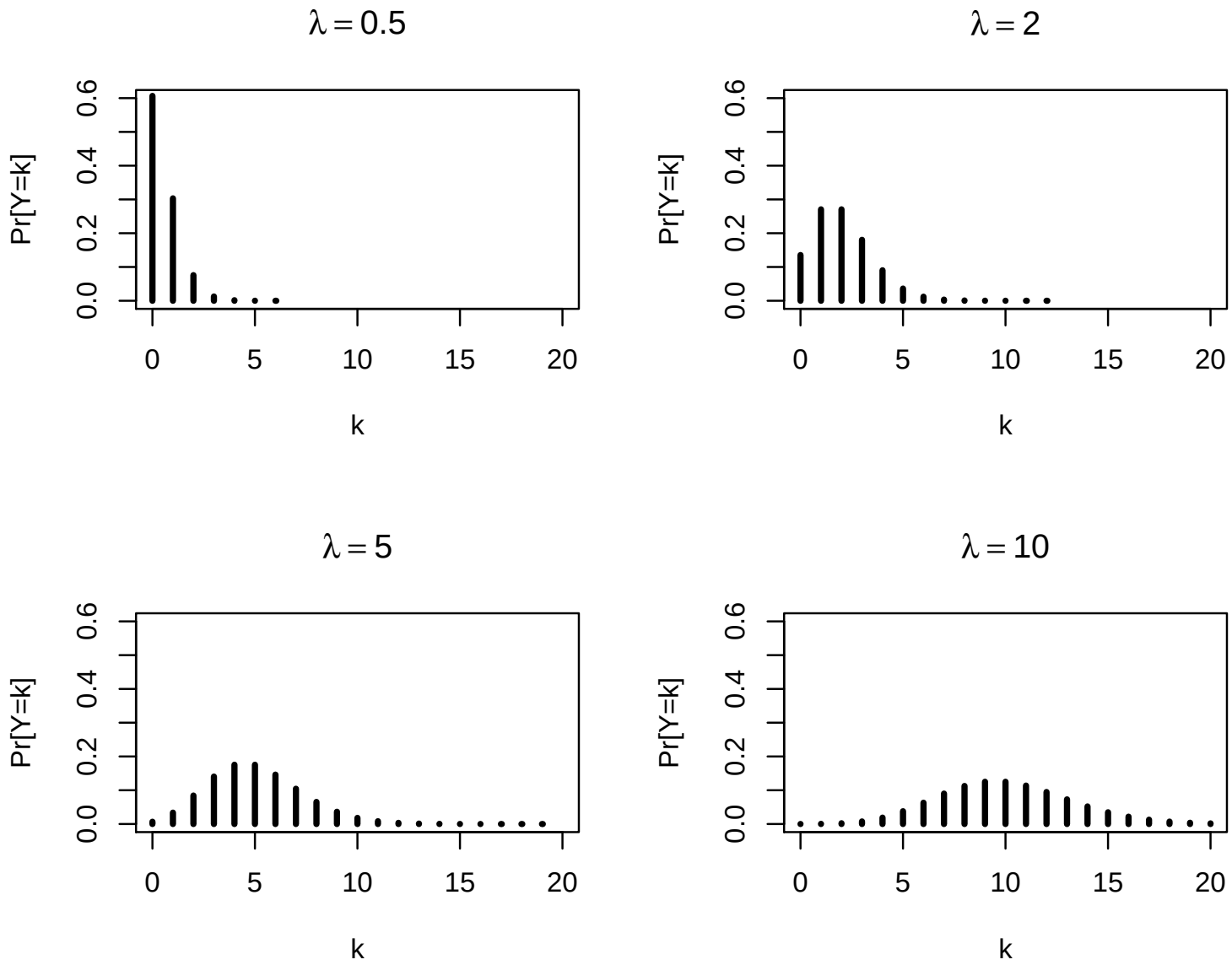
- The Poisson distribution is characterized by one parameter, λ
- If $Y \sim \text{Poisson}(\lambda)$, the probability mass function of Y is

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

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

$$E(Y) = \text{Var}(Y) = \lambda$$

Poisson Probability Mass Function



The Poisson and Binomial Distributions

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

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

i.e.

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

- Rule of thumb: $\pi \leq 0.1$ and $N \geq 20$

The Poisson and Binomial Distributions

- Table 6.6 of the 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
⋮	⋮	⋮	⋮	⋮	⋮

The Poisson and Binomial Distributions

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

The Poisson and Binomial Distributions

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

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

- 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) is approximately

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

Exact Confidence Intervals

- Cf. Note 6.8 of the text (page 195)
- Given y occurrences,

$$\hat{\lambda} = y$$

and an exact $100(1-\alpha)\%$ CI for λ is

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

Normal Approximations

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

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

- Thus an approximate $100(1-\alpha)\%$ CI for λ is

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

- A better approximation arises from

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

- For $\lambda \geq 30$, an approximate CI for $\sqrt{\lambda}$ is

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

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 a $100(1-\alpha)\%$ CI for $N\lambda$, then $(L/N, U/N)$ is a $100(1-\alpha)\%$ CI for λ. For example,

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

Example 6.20

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

$$\left[\tfrac{1}{2}\,\chi^2_{2\times 426;\,0.025},\ \tfrac{1}{2}\,\chi^2_{2\times 427;\,0.975}\right] = [386.50,\ 468.44]$$

- Normal approximations:

$$426 \pm z_{0.975} \times \sqrt{426} = [385.55, 466.45]$$

$$\left[\left(\sqrt{426} - \frac{z_{0.975}}{2}\right)^2,\ \left(\sqrt{426} + \frac{z_{0.975}}{2}\right)^2\right] = [386.51, 467.41]$$

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

Rules of Thumb

- For $\alpha = 0.05$, an approximate 95% CI for $\sqrt{\lambda}$ is

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

  implying an approximate 95% CI for λ is

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

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

$$[0, \tfrac{1}{2}\chi^2_{2;\,0.95}] \approx [0, 3.00]$$

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

Homogeneity Test

- Often, observed counts exhibit larger variance than expected under the Poisson model; this is referred to as *over-dispersion*
- This may occur if the assumption of homogeneity of the λs is not satisfied
- Want to test

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

Homogeneity Test

- Construct a χ^2 goodness of fit 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

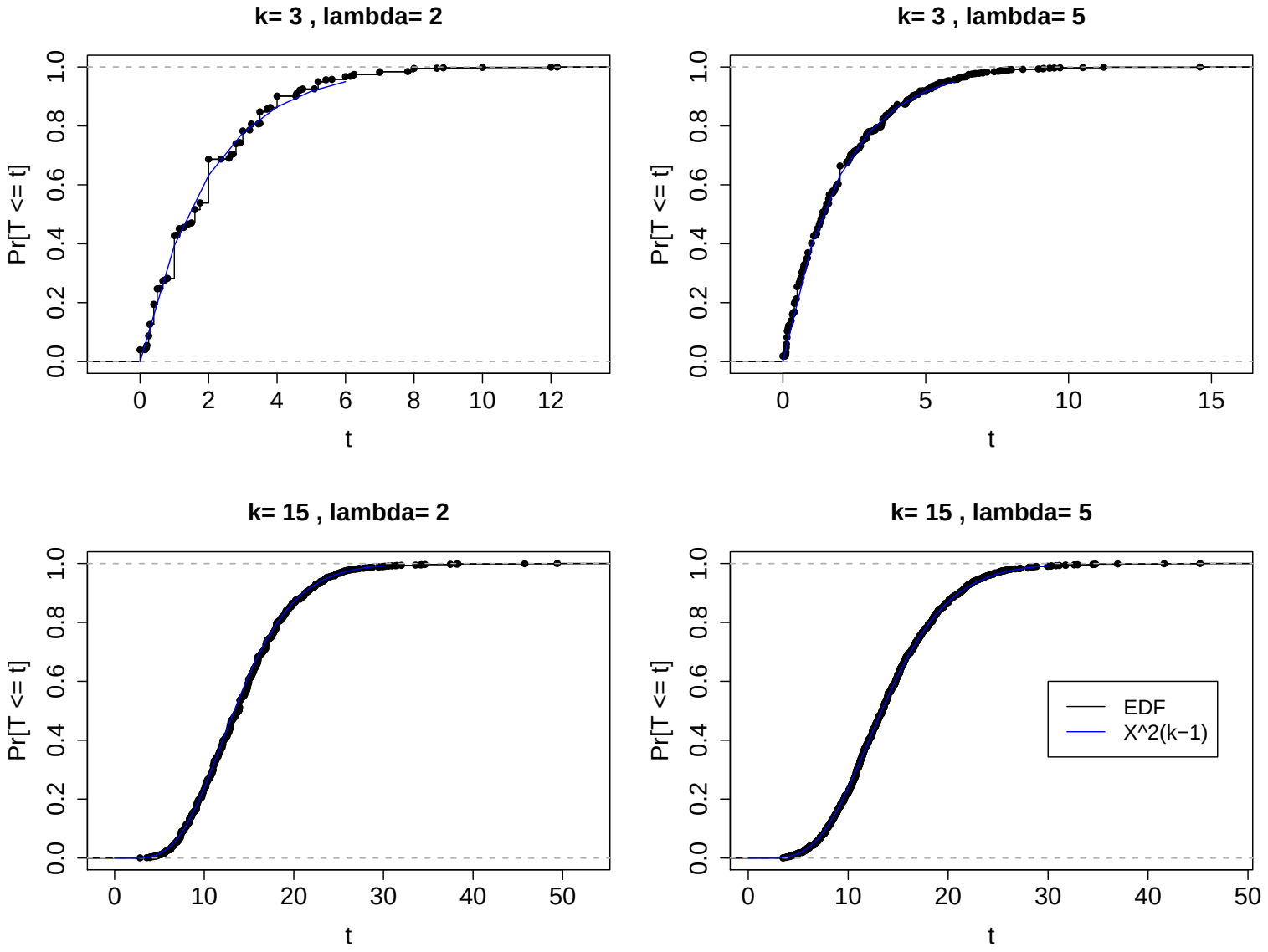
$$T = \frac{(k-1)s^2}{\bar{X}}$$

- *Poisson homogeneity/heterogeneity/dispersion test*

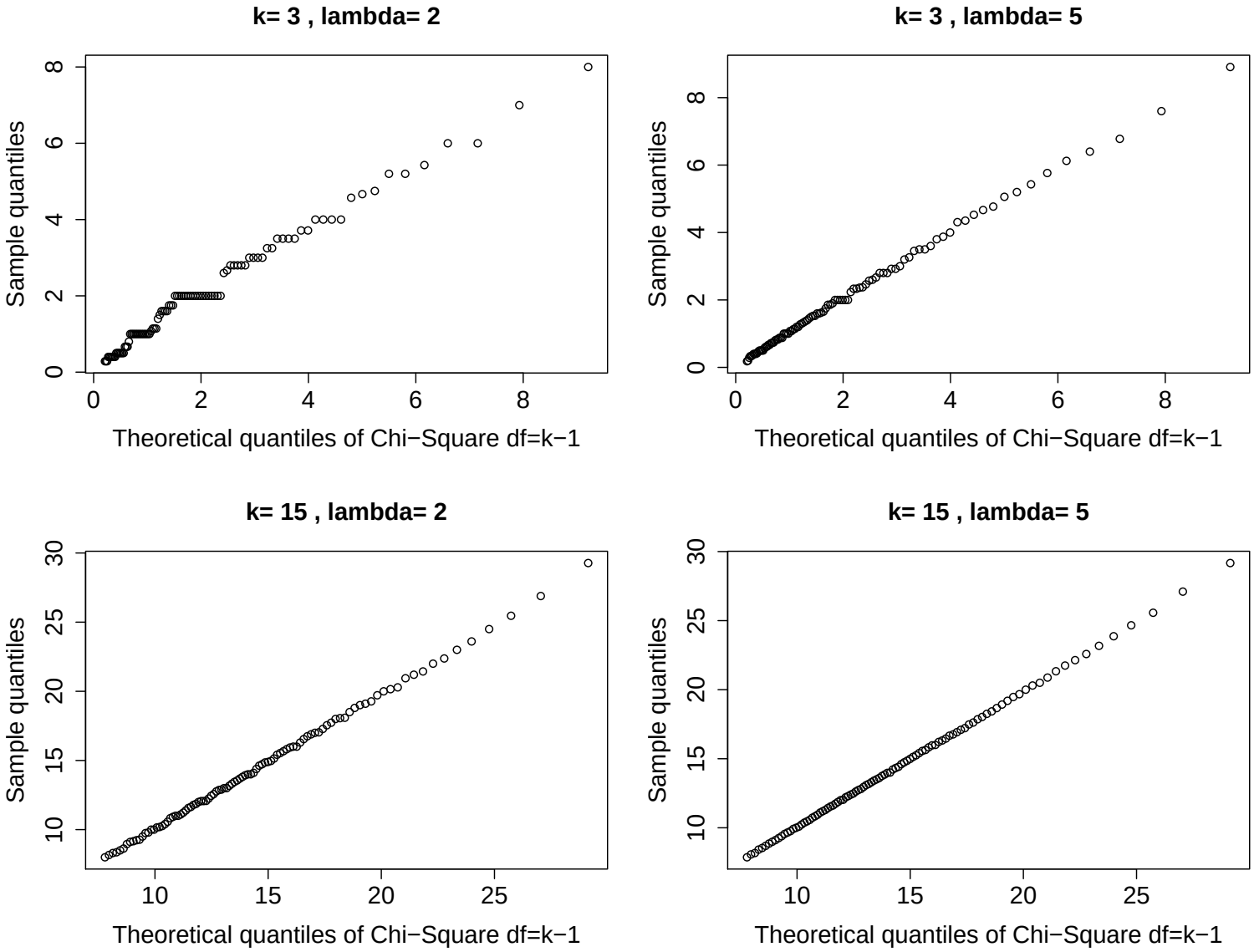
Homogeneity Test

- The χ^2 approximation 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

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

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

Negative-Binomial Distribution

- Over-dispersion may be due to heterogeneity of λs
- That is, λ is no longer a constant, but a random variable
- If λ follows a gamma distribution, then the counts follow a negative binomial distribution
- This allows for the variance to be proportional to the mean