Primer on Negative Binomial in the context of RNAseq analysis

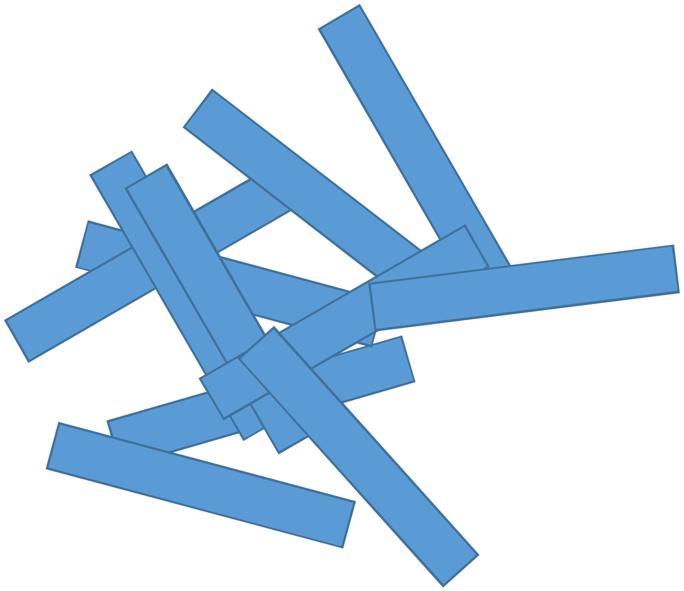
Albert Lee
Columbia University
Rabadan Lab

Outline

- Overview on analyzing count data
- Binomial -> Poisson -> Negative binomial (NB)
- Motivation for NB in RNA-seq analysis
- Derivation of NB from the marginalization of Gamma-Poisson mixture
- Mean / Variance relationship of Negative Binomial
- Why this framework works well

n flips

θ chance to see a head



n cDNA fragments (reads)

θ_g chance to map to the gene g

Let's define

X = the number of reads mapped to gene g out of n total reads, where each read has the probability $\boldsymbol{\theta_g}$

then

Let's define

X = the number of reads mapped to gene g out of n total reads,
where each read has the probability $\boldsymbol{\theta_g}$
then

$$X \sim Bin(n, \theta_g)$$

$$P(X = x) = \binom{n}{x} \theta_g^x (1 - \theta_g)^{n-x}$$

But what if you don't know n?
(total number of reads)

When $n \to \infty$

$$Bin(n, \theta) \to Pois(\lambda)$$

Where $\lambda = n\theta$

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

Proof

$$P(X = x) = \binom{n}{x}\theta^x(1-\theta)^x$$

Define $\lambda = n\theta$

$$\Rightarrow \theta = \frac{\lambda}{n}$$

$$= \frac{n!}{(n-x)!\,x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n(n-1)\ldots(n-x+1)(n-x)!}{(n-x)!\,x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n(n-1)\ldots(n-x+1)}{x!}\left(\frac{\lambda^x}{n^x}\right)\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n^x\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x}{n}+\frac{1}{n}\right)}{x!}\left(\frac{\lambda^x}{n^x}\right)\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n^x\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x}{n}+\frac{1}{n}\right)}{x!}\left(\frac{\lambda^x}{n^x}\right)\left(1-\frac{\lambda}{n}\right)^{n-x}$$

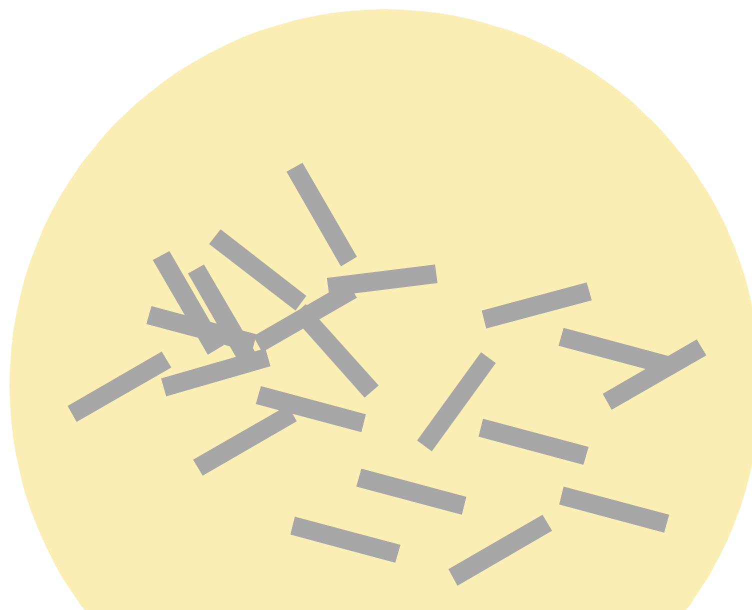
$$= \frac{\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x}{n}+\frac{1}{n}\right)}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

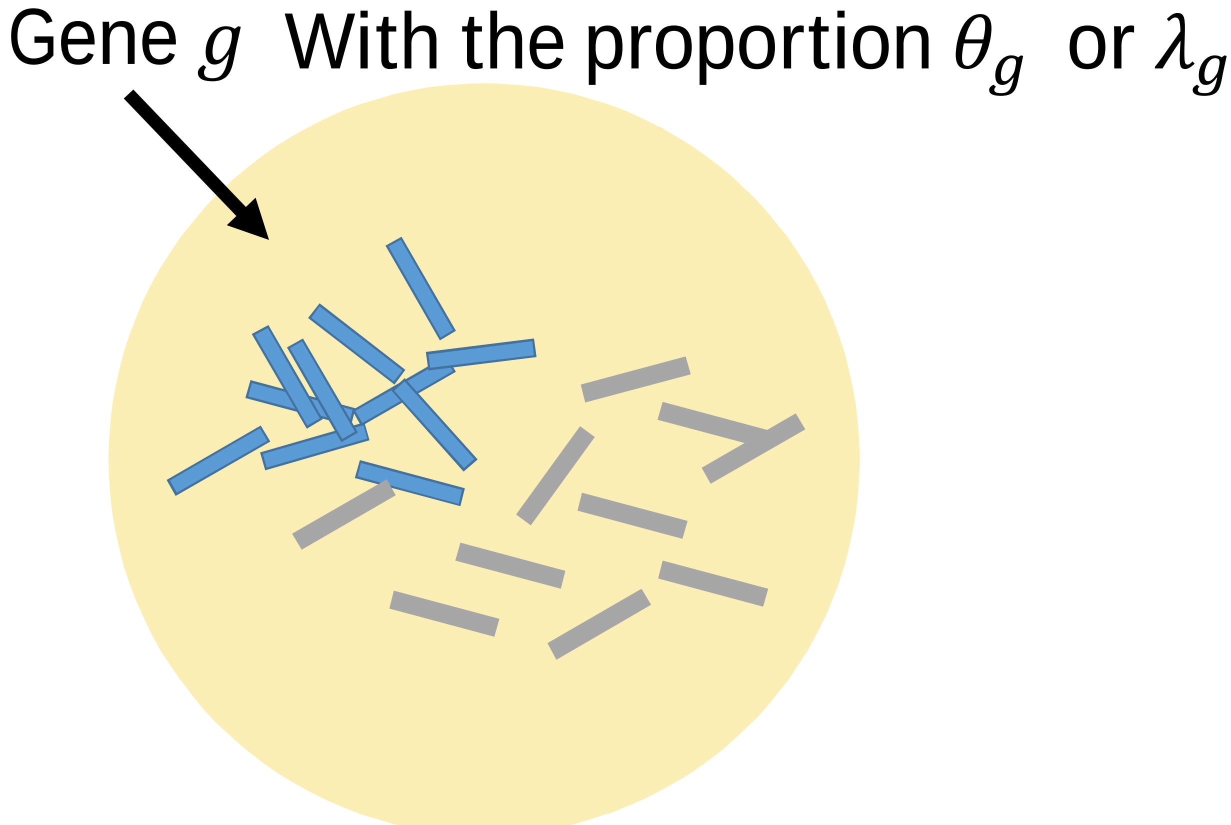
$$= \frac{\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x}{n}+\frac{1}{n}\right)}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x}$$

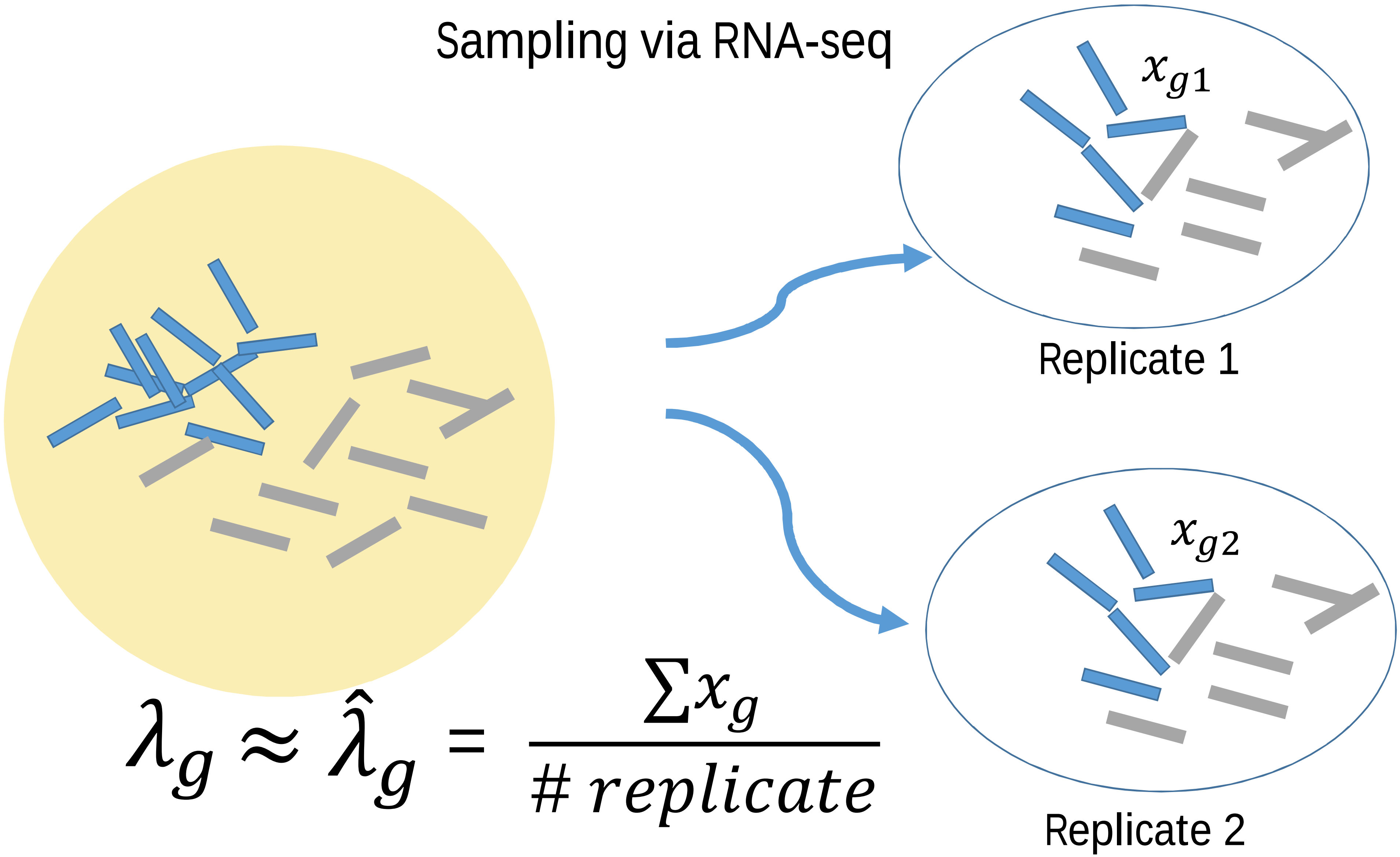
$$n \to \infty$$

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$$

$$= \frac{1}{x!}\lambda^x e^{-\lambda}$$

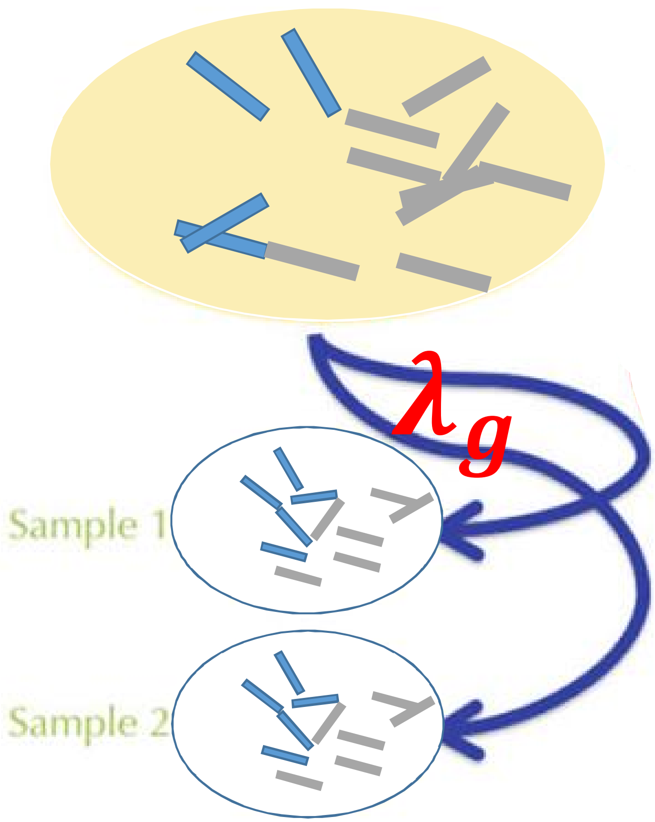



Gene g With the proportion θ_g or λ_g



$$\lambda_g \approx \hat{\lambda}_g = \frac{\sum x_g}{\#\, replicate}$$

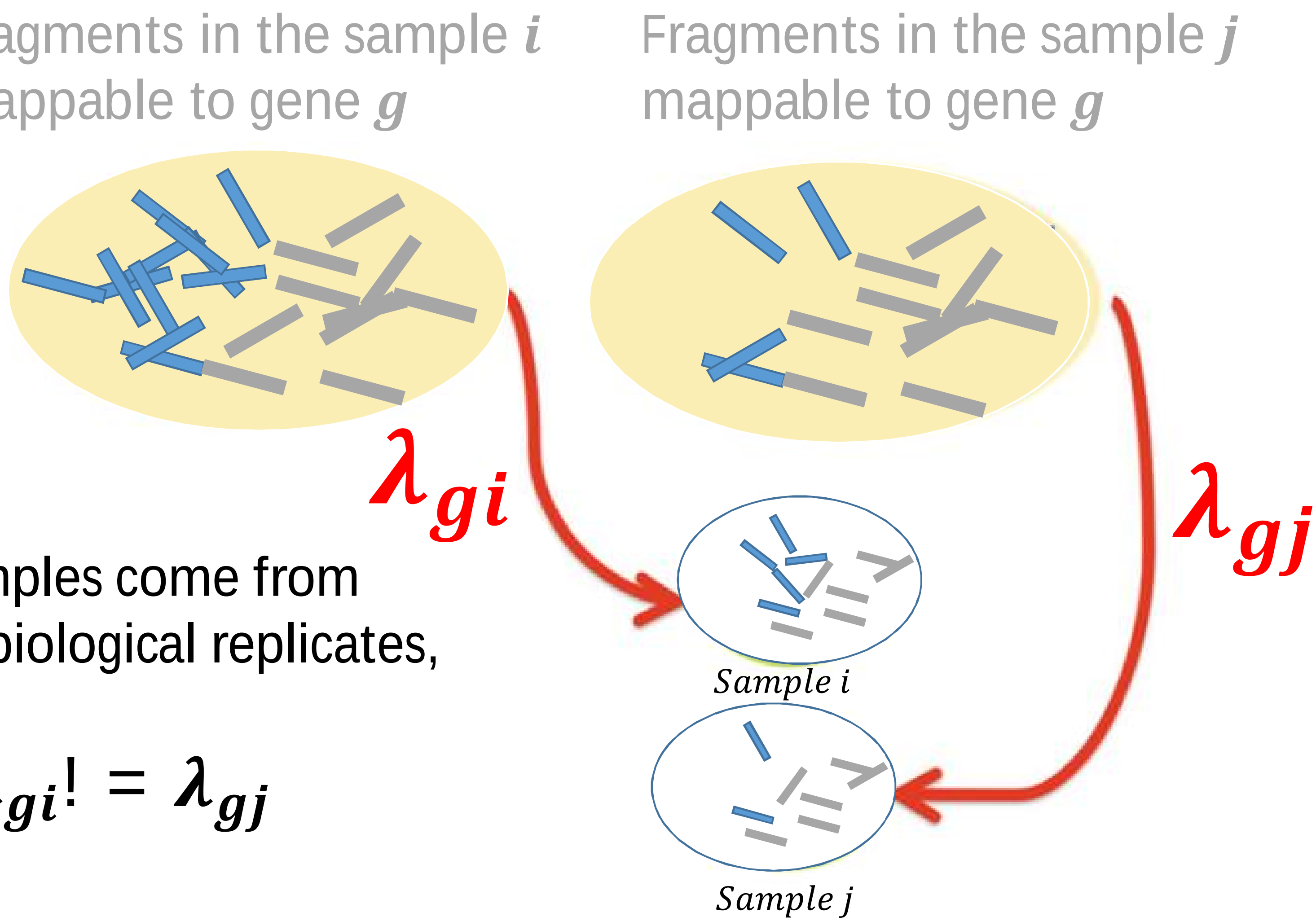
cDNA in the sample $\boldsymbol{i}$ mappable to gene $\boldsymbol{g}$



But this only reflects when $\boldsymbol{\theta_g}$ (or $\boldsymbol{\lambda_g}$)is constant

=> Technical replicates

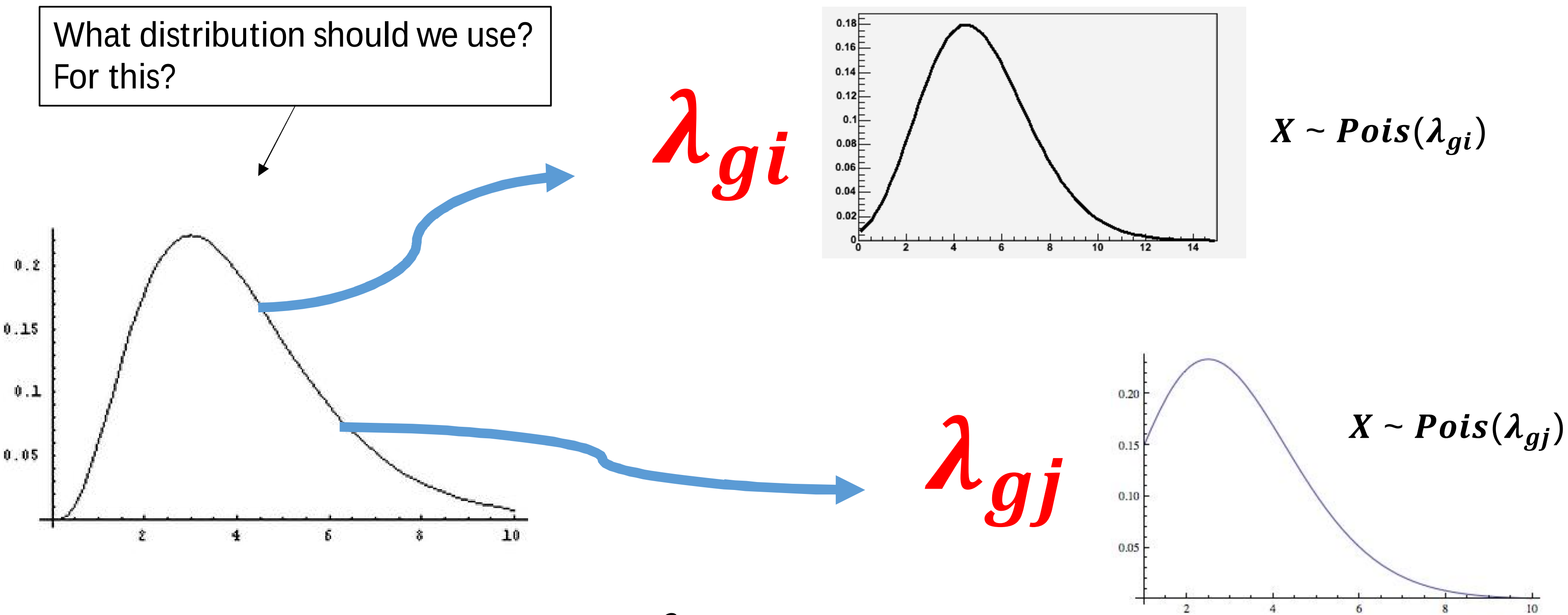
But..



When samples come from
Different biological replicates,

$$\lambda_{gi}! = \lambda_{gj}$$

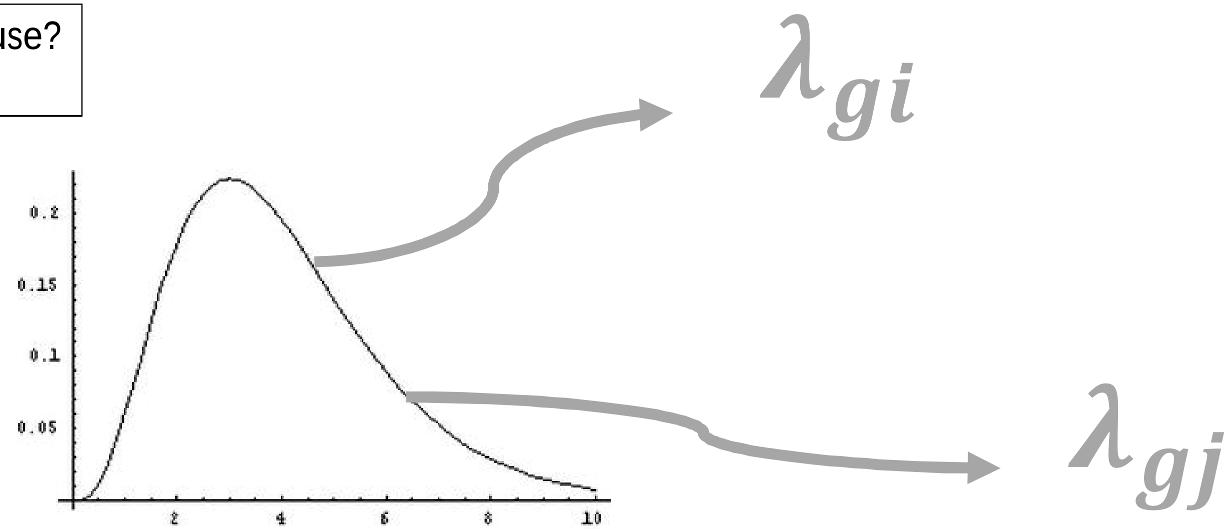
What do you do?



Make λ a random variable!

We use gamma distribution

What distribution should we use?
For this?



$$P_{gamma}(X = x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

Why gamma?

- It's mathematically convenient

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

Shape parameter $\boldsymbol{\alpha}$

Rate parameter $\boldsymbol{\beta}$

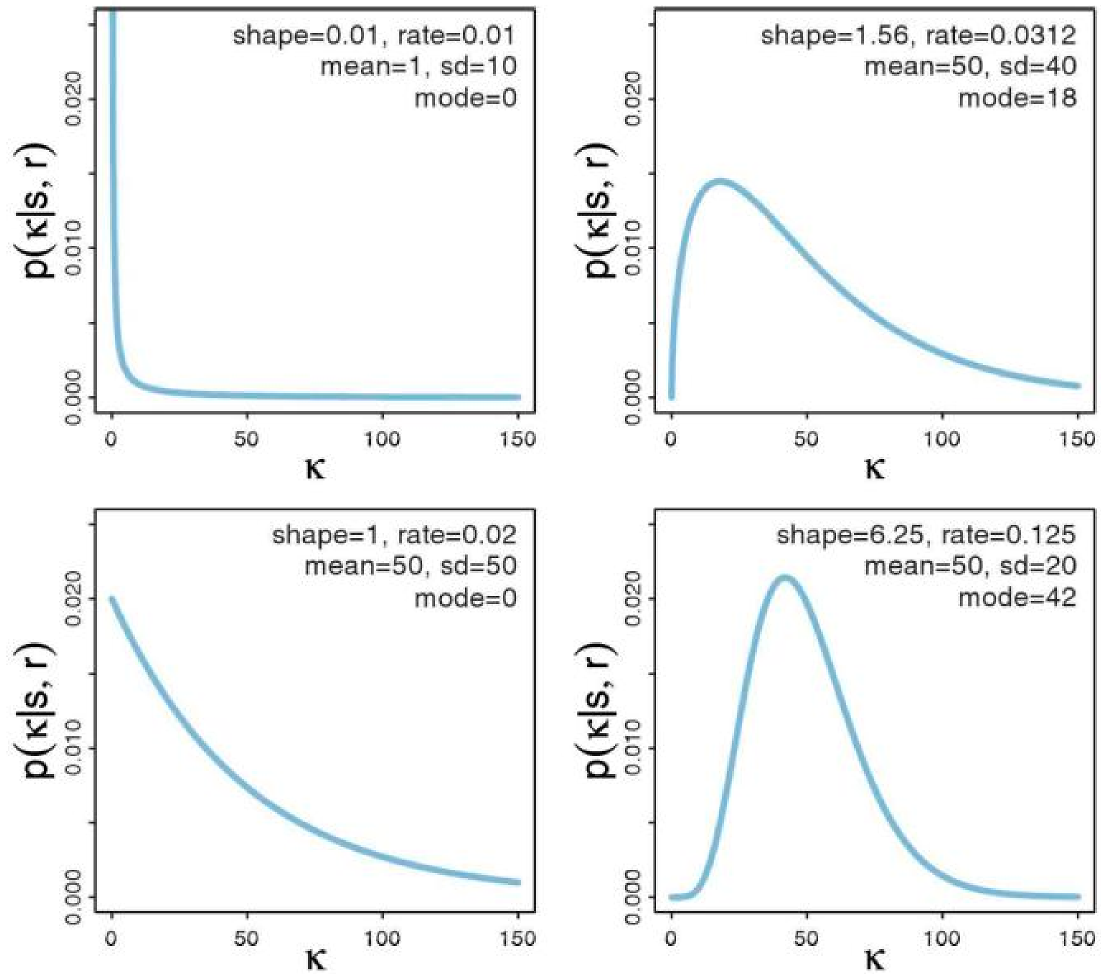
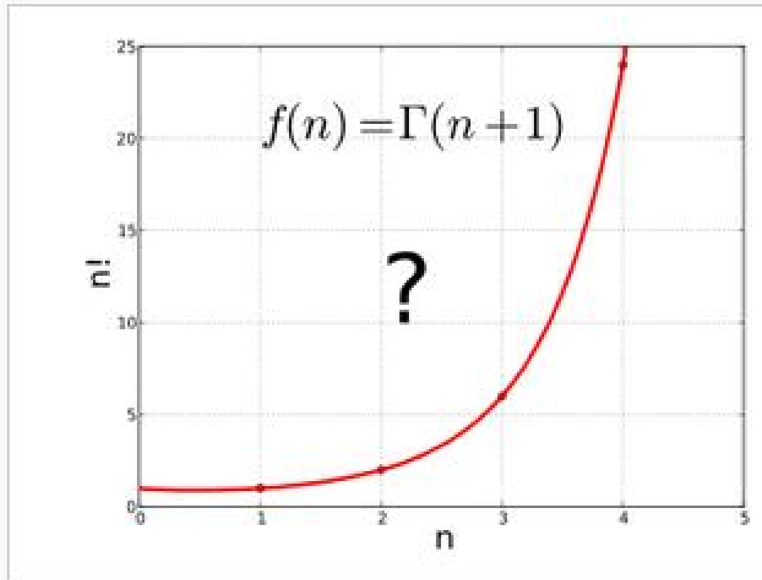


Figure 9.8 Examples of the gamma distribution. The vertical axis is $p(\kappa|s,r)$ where s is the shape and r is the rate, whose values are annotated in each panel. From Doing Bayesian Analysis 2nd ed

What is Gamma function? $\Gamma(t)$ in $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$

$\Gamma(t) = (t-1)!$ if $t \in \mathbb{Z}$ & $t - 1 \geq 0$

but what if $t \in \mathbb{R}$ & $t - 1 \geq 0$

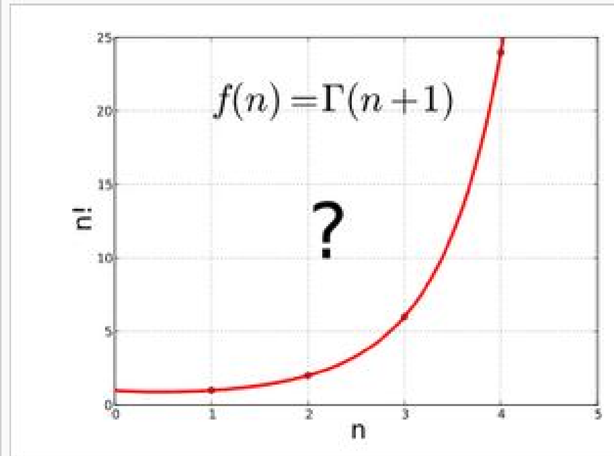


It is easy graphically to interpolate the factorial function to non-integer values, but is there a formula that describes the resulting curve?

What is Gamma function? $\Gamma(t)$

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$

$$\frac{\Gamma(t)}{a^t} = \int_0^{\infty} x^{t-1} e^{-ax} dx$$



It is easy graphically to interpolate the factorial function to non-integer values, but is there a formula that describes the resulting curve?

$$\int_0^{\infty} x^{t-1} e^{-(ax)} dx = a^{t-1} a^{1-t} \int_0^{\infty} x^{t-1} e^{-(ax)} dx$$

$$= a^{1-t} \int_0^{\infty} (ax)^{t-1} e^{-(ax)} dx$$

let $u = ax$
$du = a\, dx$

$$= a^{-t} \int_0^{\infty} u^{t-1} e^{-u} du = a^{-t}\Gamma(t)$$

Why Gamma distribution is valid prob. Dist?

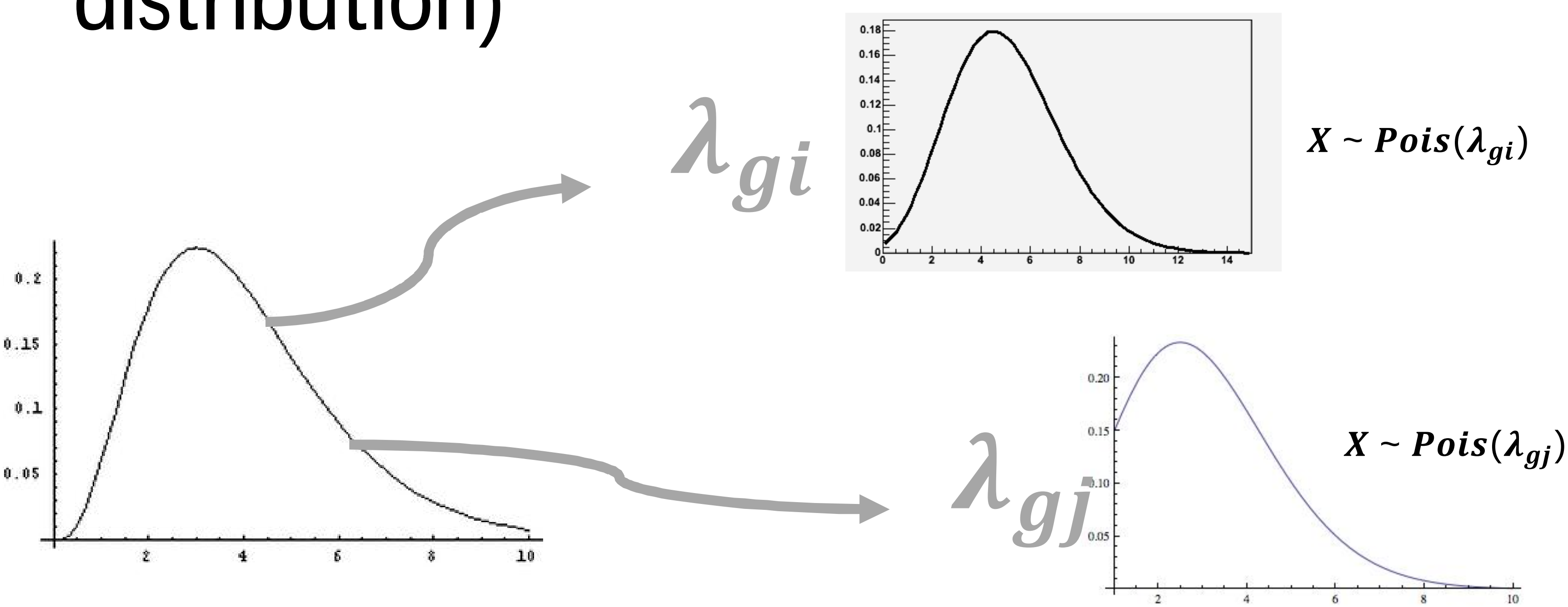
$$\mathrm{P}_{gamma}(X = x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

$$\int \mathrm{P}_{gamma}(X = x) = \int \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)\, dx$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int x^{\alpha-1} \exp(-\beta x)\, dx$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \beta^{-\alpha} \Gamma(\alpha) = 1$$

Hierarchical modeling (or Mixture distribution)



$$\Lambda \sim \Gamma\,(a, \beta)$$
$$X \mid \Lambda = \lambda \sim Pois(\lambda)$$

Gamma-Poisson mixture

$$\Lambda \sim \Gamma(\alpha, \beta) \quad \rightarrow P(\lambda)$$

$$(X \mid \Lambda = \lambda) \sim Pois(\lambda) \quad \rightarrow P(x \mid \lambda)$$

$$P(X = x) = ?$$

$$P(X = x) = \int_0^\infty P(\lambda, x)\, d\lambda$$

marginalization

$$= \int_0^\infty P(x \mid \lambda) P(\lambda) d\lambda$$

Chain rule

$$= \int_0^\infty \left(\frac{\lambda^x e^{-\lambda}}{x!}\right)\left(\frac{\beta^\alpha}{\Gamma(a)} \lambda^{a-1} e^{(-\beta\lambda)}\right) d\lambda$$

$$= \int_0^\infty \left(\frac{\lambda^x e^{-\lambda}}{x!}\right)\left(\frac{\beta^\alpha}{\Gamma(a)}\lambda^{a-1}e^{-\beta\lambda}\right)d\lambda$$

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\int_0^\infty \lambda^x e^{-\lambda}\lambda^{a-1}e^{-\beta\lambda}d\lambda$$

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\int_0^\infty \lambda^{(x+a)-1}e^{-(\beta+1)\lambda}d\lambda$$

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\left(\frac{(\beta+1)^{x+\alpha}}{\Gamma(x+a)}\right)^{-1}\left(\frac{(\beta+1)^{x+\alpha}}{\Gamma(x+a)}\right)\int_0^\infty \lambda^{(x+a)-1}e^{-(\beta+1)\lambda}d\lambda$$

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\left(\frac{(\beta+1)^{x+\alpha}}{\Gamma(x+a)}\right)^{-1}\left(\frac{(\beta+1)^{x+\alpha}}{\Gamma(x+a)}\right)\int_0^\infty \lambda^{(x+a)-1}e^{-(\beta+1)\lambda}d\lambda$$

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\left(\frac{(\beta+1)^{x+\alpha}}{\Gamma(x+a)}\right)^{-1}(1)$$

This is negative binomial!

$$= \frac{1}{x!}\frac{\beta^\alpha}{\Gamma(a)}\left(\frac{\Gamma(x+a)}{(\beta+1)^{x+\alpha}}\right) = \frac{\Gamma(x+a)}{x!\,\Gamma(a)}\left(\frac{1}{\beta+1}\right)^x\left(\frac{\beta}{\beta+1}\right)^a$$

$$= \frac{\Gamma(x+a)}{x!\,\Gamma(a)} \left(\frac{1}{\beta+1}\right)^x \left(\frac{\beta}{\beta+1}\right)^a$$

Let's restrict that x+a is integer, then $\Gamma(x) = (x-1)!$

$$= \left(\frac{(x+a-1)!}{x!\,(a-1)!}\right) \left(\frac{1}{\beta+1}\right)^x \left(\frac{\beta}{\beta+1}\right)^a$$

Because $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

$$= \binom{x+a-1}{a-1} \left(\frac{1}{\beta+1}\right)^x \left(\frac{\beta}{\beta+1}\right)^a$$

Define $p = \frac{1}{\beta+1}$ $\Rightarrow 1-p = \frac{\beta}{\beta+1}$

$$= \binom{x+a-1}{a-1} p^x (1-p)^a$$

If X is the number of success before a failures, then

$$X \sim NB(a, p)$$

Why ?

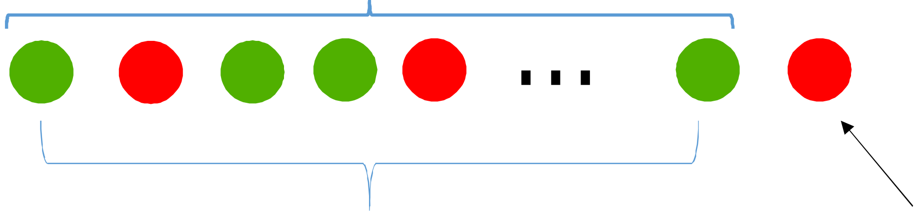
If X is the number of success before a failures, then

$$X \sim NB(a, p)$$

X= number of success

a= number of failures

But the sequence of the balls before the last step can very



$(x + a - 1)$ *many balls*

By definition, failure must happen at the end

$\binom{x+a-1}{a-1}$= total number of possible combination of failures before the last step

$$= \binom{x+a-1}{a-1} p^x (1-p)^{a-1} (1-p)$$

$$= \binom{x+a-1}{a-1} p^x (1-p)^a$$

We know that for Poisson :

$$E\left[X_{pois}\right] = \lambda$$

$$Var\left[X_{pois}\right] = \lambda$$

But for NB, what's

$$E[X_{NB}] = ?$$

$$Var[X_{NB}] = ?$$

Mean of negative binomial

$$\mathrm{E}[X] = \int x P(X) dx$$

$$= \int_0^\infty x \frac{\Gamma(x+r)}{x!\,\Gamma(r)} p^x (1-p)^r dx$$

NASTY !

Instead solve it via Moment generating function of NB

Moment generating function = magical formula that spits out things we need:
mean,variance,skewness, kurtosis,…
or what's collectively known as
"moments"(shape of the distribution)

$$M_X(t) = E[e^{tX}]$$

$$M_X{}'(0) = E[X]$$ Mean

$$M_X{}''(0) = E[X^2]$$

$$Var[X] = E[X^2] - E[X]^2$$ Variance

"Secret of the magic trick"

$$e^{x} = \sum_{k=0}^{\infty} \frac{x^{k}}{k!} = \frac{x^{0}}{0!} + \frac{x^{1}}{1!} + \frac{x^{2}}{2!} + \frac{x^{3}}{3!} + \ldots$$

$$e^{tX} = \frac{tX^{0}}{0!} + \frac{(tX)^{1}}{1!} + \frac{(tX)^{2}}{2!} + \frac{(tX)^{3}}{3!} + \ldots$$

$$M_X(t) = \mathrm{E}[e^{tX}] = \mathrm{E}\left[\frac{(tX)^{0}}{0!} + \frac{(tX)^{1}}{1!} + \frac{(tX)^{2}}{2!} + \frac{(tX)^{3}}{3!} + \ldots\right]$$

$$= \mathrm{E}\left[\frac{(tX)^{0}}{0!}\right] + E\left[\frac{(tX)^{1}}{1!}\right] + E\left[\frac{(tX)^{2}}{2!}\right] + E\left[\frac{(tX)^{3}}{3!}\right] + \ldots$$

Linearity of expectation

$$= 1 + tE[X] + \frac{t^{2}E[X^{2}]}{2!} + \frac{t^{3}E[X^{3}]}{3!} + \ldots$$

E[cX] = cE[X]

$$M_X'(t) = 0 + E[X] + tE[X^{2}] + \frac{t^{2}E[X^{3}]}{2!} + \cdots \longrightarrow M_X'(0) = E[X]$$

$$M_X''(t) = E[X^{2}] + tE[X^{3}] + \cdots \longrightarrow M_X''(0) = E[X^{2}]$$

Instead solve it via
Moment generating function of NB

$$M_X(t) = E[e^{tX}]$$

$$= \sum_{x=0}^{\infty} e^{tx} \binom{x+r-1}{x} p^x (1-p)^a$$

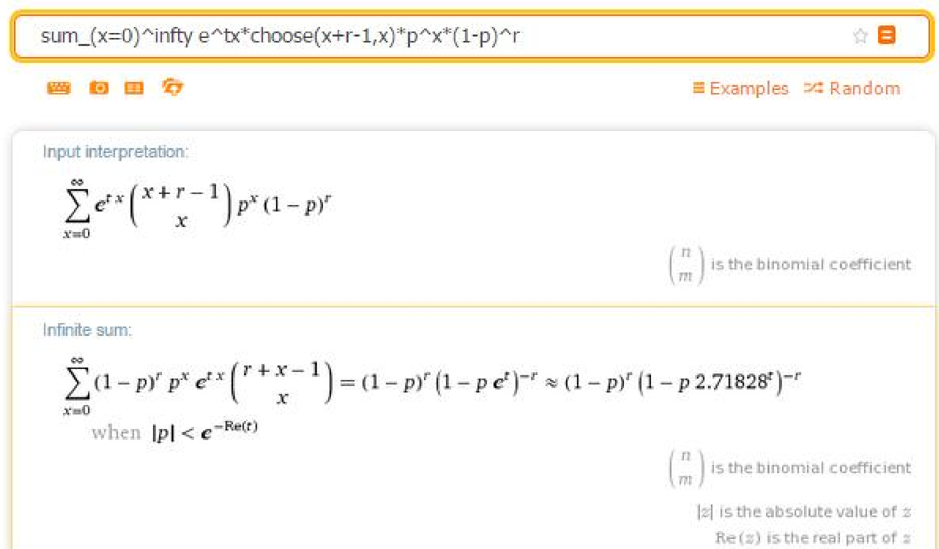
$$= (1-p)^a \sum_{x=0}^{\infty} e^{tx} \binom{x+r-1}{x} p^x$$

$$= (1-p)^a \sum_{x=0}^{\infty} \binom{x+r-1}{x} (pe^t)^x$$

$$= \frac{(1-p)^a}{(1-pe^t)^a}$$

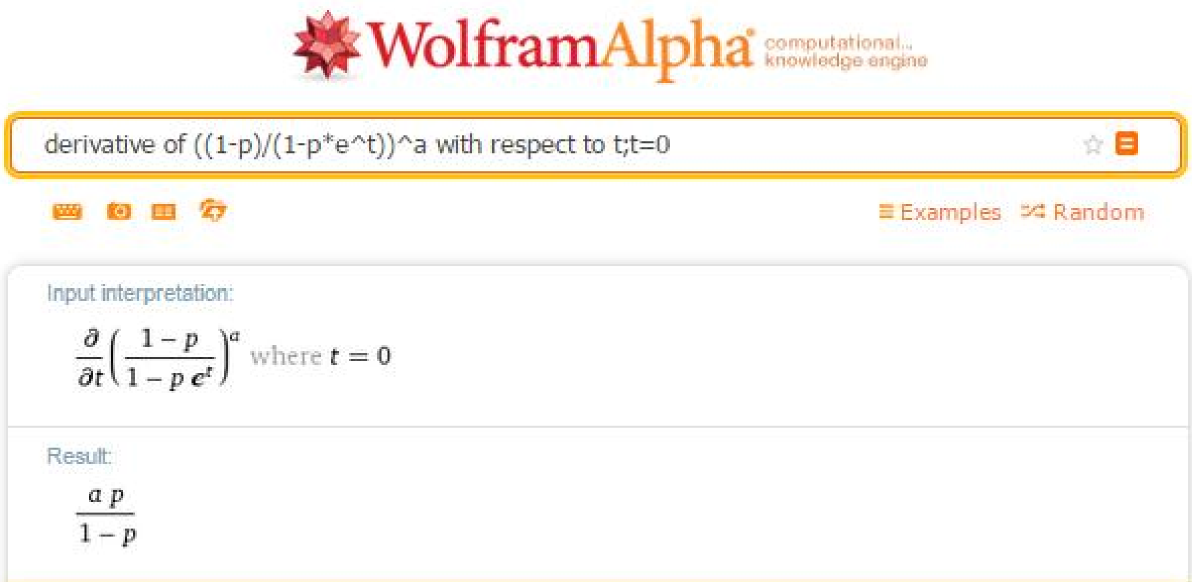
$$M_{X_{NB}}(t) = \left(\frac{1-p}{1-pe^t}\right)^r$$

$$\boxed{\sum_{x=0}^{\infty} \binom{x+a-1}{x} (pe^t)^x = \frac{1}{(1-pe^t)^a}}$$



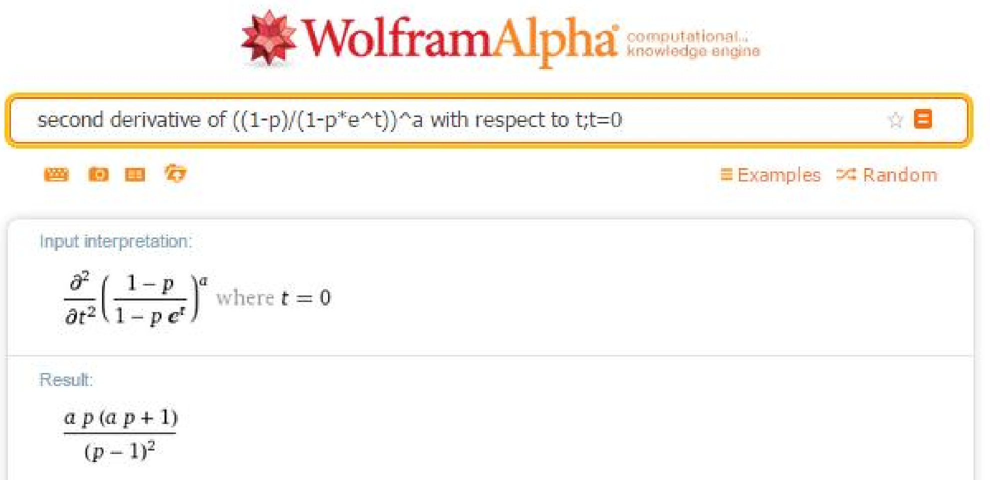
Mean of NB

$$M_X(t) = \left(\frac{1-p}{1-pe^t}\right)^r$$



$$M'_X(0) = \frac{ap}{1-p} = E[X]$$

Variance of NB



$$M_X''(0) = \frac{ap(ap+1)}{(p-1)^2} = E[X^2]$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{ap(ap+1)}{(p-1)^2} - \left(\frac{ap}{1-p}\right)^2 = \frac{ap}{(1-p)^2}$$

Denouement

We have $E[X] = \frac{ap}{1-p}$ $Var[X] = \frac{ap}{(1-p)^2}$

Remember $p = \frac{1}{\beta + 1}$ $1 - p = \frac{\beta}{\beta + 1}$

Then

$$E[X] = \frac{ap}{1-p} = \frac{a\left(\frac{1}{\beta+1}\right)}{\frac{\beta}{\beta+1}} = \frac{a}{\beta}$$

$$Var[X] = \frac{ap}{(1-p)^2} = \frac{a\left(\frac{1}{\beta+1}\right)}{\left(\frac{\beta}{\beta+1}\right)^2} = \frac{a}{\beta^2}(\beta + 1)$$

$$= \frac{a}{\beta} + \frac{a}{\beta^2}$$

$$= \frac{a}{\beta} + \left(\frac{a}{\beta}\right)^2\left(\frac{1}{a}\right)$$

Therefore

$$Var[X] = E[X] + \left(\frac{1}{a}\right)E[X]^2$$

We know that for poisson :

$$E[X_{pois}] = \lambda$$
$$Var[X_{pois}] = \lambda$$

And for NB,

$$E[X_{NB}] = \mu$$
$$Var[X_{NB}] = \mu + c\mu^2$$

Dispersion

Most efforts of most RNAseq packages...

$$E[X_{NB}] = \mu$$
$$Var[X_{NB}] = \mu + c\mu^2$$

Estimate dispersion using small # of samples

EdgeR : (old) assume c is identical for all genes

(new) c is different for individual genes

Using hierarchical modeling to shrink to c to a consensus value

DESeq : Assume c and μ is related nonlinearly. use local regression to compute c

Why is this nice???

- Negative binomial is exponential family! So you can model complicated stuff using generalized linear model.

$$E[X_{NB}] = \mu$$

Assume $\mu = e^{\mathrm{X}^{\mathrm{T}}\beta}$

Estimate β

X = design matrix

β= vector of coefficients

- Empirically, packages that use NB as an underlying model are shown to work better than those that don’t

THE END