High-Throughput Sequencing Course

Count Models for RNA-Seq

Biostatistics and Bioinformatics

Summer 2018

Duke University School of Medicine

DUKE BIOSTATISTICS & BIOINFORMATICS

Two Approaches for Analysis of RNA-Seq

- Two-stage method: Convert counts to "Expression" and then use statistical methods for microarrays (e.g., t-test)
- One-stage method: Relate the counts directly to the phenotype
- This is done through using statistical methods for modeling counts
- We generally promote the latter approach for data analysis

DESeq for RNA-Seq

- The goal is to provide sufficient background to understand the DESeq method
- We are not suggesting that DESeq is the best approach for analysis of RNA-Seq data
- We are considering it in this course as one, of many other methods, that adhere to the one-stage approach principle
- Added bonus: Nicely written `R` extension package (important feature for teaching)
- DESeq has many limitations (e.g., it cannot directly deal with quantitative and censored outcomes)
- Also some of the theoretical details (e.g., the effect of using plugin estimates for nuisance parameters) have seemingly not been fully fleshed out

Three Distributions for Count Data

- RNA-Seq data are counts (not continuous measurements)
- To properly model RNA-Seq data, we need to consider distributions to model counts
- We will consider three important distributions for counts:
  - Binomial
  - Poisson
  - Negative Binomial
- There are many other distributions for counts (e.g., geometric distribution) that will not be discussed
- Brief notes on multinomial distribution

Distribution for Counts: Support

- A count is a non-negative (zero or positive) integer
- When considering a distribution of a count variable, we first have to determine its *support*
- The support of the distribution consists of the values that could occur with positive probability
- For example, if we toss a coin once and we count the number of heads, the support is $\{0, 1\}$
- If we flip it twice, the support is $\{0, 1, 2\}$
- Why is 3 not in the support? How about -1?
- These values are not *possible* (they have zero probability)
- The probability to observed three heads among two tosses is zero.

Distribution for Counts: Probability Mass Function

- Example: we toss a fair coin once and we count the number of heads (call it K)

$$P(K = 0) = \frac{1}{2} \text{ and } P(K = 1) = \frac{1}{2}$$

  and

$$P(K = k) = 0$$

  if k is not 0 or 1
- The probability mass function (PMF) determines the probability that K assumes value k in the support
- Sometimes we use the terms "distribution" and "PMF" interchangeably

Distribution for Counts: Probability Mass Function

- Example: we toss a fair coin twice and we count the number of heads (call it K)

$$P(K = 0) = \frac{1}{4} \text{ and } P(K = 1) = \frac{1}{2} \text{ and } P(K = 2) = \frac{1}{4}$$

- Why?
- Note that if once adds up $P(K = k)$ over all k in the support the sum should be one

$$\sum_k P(K = k) = 1$$

Exercise: Support and PMF

- we toss a biased coin twice and we count the number of heads (call it K)
- the probability that any toss lands a head is $\pi = \frac{1}{3}$
- What is the support of the distribution
- What is the PMF
- Repeat the last steps if π is any arbitrary number (between 0 and 1 of course)

Exercise: Support and PMF

- the support is as in the previous example $\{0, 1, 2\}$
- Why is it unchanged

$$P(K = 0) = \frac{4}{9} \text{ and } P(K = 1) = \frac{4}{9} \text{ and } P(K = 2) = \frac{1}{9}$$

- More generally

$$\begin{aligned} P(K = 0) &= (1 - \pi)^2 \\ P(K = 1) &= 2\pi(1 - \pi) \\ P(K = 2) &= \pi^2 \end{aligned}$$

Flipping the coin

- Throughout this discussing we will consider flipping a coin
- The coin lands a head with probability π (could be biased) or tail with probability $1 - \pi$
- For convenience, we will recode H as 1 and T as 0
- We will flip it n times.
- Notation:
  - n is to denote the number of *trials*
  - On any trial (or flip), if we land an H we will call it an event (or success)
  - or if we land a T we will call it a failure
- RNA-seq connection: You can think of a read mapping to a gene to be an event

Three Variants of the Coin Tossing Experiment

1. Fix the number of trials (n) upfront and then toss the coin n times
   - The number of events (among n trials) is random
2. Toss the coin a large number of times and assume that each one of these many trials has a small probability of being an event
   - Here n is large and π is small (close to 0)
3. Fix the number of desired events upfront, then toss the coin repeatedly to achieve that number
   - Here the number of trials n is random

Outline for today (and maybe Thur)

- Provide an overview of the properties of the three distributions
- PMF, Mean and Variance
- Discuss relationship between the three distributions
- Need to introduce some notation (unfortunately)
- The goal is develop a regression model for counts
- We motivate this first using linear regression
- And then through logistic regression
- Before moving on to dicussing a regression model based on negative binomial distribution
- Provide some insight on how these models are estimated

Example: Fixed n

- We flip the coin $n = 6$ times
- Observed sequence: TTHTTH
- We recode this as 001001
- This corresponds to
  - $n = 6$ trials
  - 2 events (or successes)
  - or equivalently 4 failures

Number of possible Outcomes

- Example 1: Suppose that $n = 2$
  - 4 possible outcomes: $\{00, 10, 01, 11\}$
  - $4 = 2 \times 2 = 2^2$
- Example 2: Suppose that $n = 3$
  - Eight possible outcomes: $\{000, 100, 101, 001, 110, 011, 101, 111\}$
  - $8 = 2 \times 2 \times = 2^3$
  - Example 3: $n = 6$
  - $64 = 2^6$ outcomes
- The number of possible outcomes based on n trials is 2^n
- But we are not interested in counting outcomes
- We want to count the number of outcomes corresponding to $K = 0, K = 1, ..., K = n$

Number of Successes

- Example 1: Suppose that $n = 2$
  - 4 possible outcomes: $\{00, 10, 01, 11\}$
  - Number outcomes corresponding to $K = 0$ is 1
  - Number outcomes corresponding to $K = 1$ is 2
  - Number outcomes corresponding to $K = 2$ is 1
- Example 2: Suppose that $n = 3$
  - Eight possible outcomes: $\{000, 100, 010, 001, 110, 011, 101, 111\}$
  - Number outcomes corresponding to $K = 0$ is 1
  - Number outcomes corresponding to $K = 1$ is 3
  - Number outcomes corresponding to $K = 2$ is 3
  - Number outcomes corresponding to $K = 3$ is 1
- What does this look like for a general n?
- If you toss the coin n times, how many outcomes correspond to k events?

Factorial Function

- Integers are "whole" numbers $\ldots, -2, -1, 0, 1, 2, \ldots$
- Consider a non-negative integer k $(0, 1, 2, \ldots)$
- $0! = 1$
- $1! = 1$
- $2! = 2 \times 1 = 2$
- $3! = 3 \times 2 \times 1 = 6$
- $4! = 4 \times 3 \times 2 = 24$
- $\cdots$
- $k! = k \times (k-1) \times (k-2) \times \ldots 3 \times 2 \times 1$

Number of Combinations

- The number of possible combinations on the basis of k events among n trials

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

- Example 1: Suppose that $n = 3$ and $k = 1$

$$\binom{3}{1} = \frac{3!}{1!(2-1)!} = \frac{3 \times 2 \times 1}{1 \times 2 \times 1} = 3$$

```
choose(3, 1)

## [1] 3
```

- Example 2: Suppose that $n = 4$ and $k = 2$

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = \frac{24}{4} = 6$$

```
choose(4, 2)

## [1] 6
```

Toss the coin n times

- Toss the coin n times
- Number of possible outcomes: 2^n
- Number outcomes corresponding to $K = 0$ is 1.
- Number outcomes corresponding to $K = 1$ is n.
- Number outcomes corresponding to $K = n - 1$ is n.
- Number outcomes corresponding to $K = n$ is 1.
- Number outcomes corresponding to $K = k$ is

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

  for $k = 0, 1, 2, \ldots$
- Do the results for $K = 0, 1, n - 1, K = n$ agree with the formula?
- Related to the Pascal Triangle

Pascal Triangle

$$
\begin{array}{lccccccccccccc}
n=0 & & & & & & & 1 & & & & & & \\
n=1 & & & & & & 1 & & 1 & & & & & \\
n=2 & & & & & 1 & & 2 & & 1 & & & & \\
n=3 & & & & 1 & & 3 & & 3 & & 1 & & & \\
n=4 & & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
n=5 & & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
n=6 & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}
$$

Bernoulli Distribution

- Suppose that we toss the coin just once
- In other words $n = 1$
- We say that the number of events follows a Bernoulli distribution with parameter π
- The PMF is

$$P(K = k) = \pi^k (1 - \pi)^{1-k}, k = 0, 1$$

```
set.seed(12324)
# Simulate 10 Bernoulli random variables with parameter pi=0.5
rbinom(10, 1, 0.5)
```

```
##  [1] 1 1 1 1 1 0 0 0 0 0
```

```
# Simulate 5 Bernoulli random variables with parameter pi=0.23
rbinom(5, 1, 0.23)
```

```
## [1] 0 0 0 0 0
```

Binomial Distribution

- For the Bernoulli distribution $n = 1$
- More generally (when $n \geq 1$) the number of events K is said to follow a Binomial distribution with parameters n and π
- The distribution is

$$P[K = k] = \binom{n}{k} \pi^k (1 - \pi)^{n-k},$$

  $k = 0, 1, 2, \ldots, n$
- Note that when $n = 1$ the Binomial reduces to a Bernoulli distribution . Why?
- Why is does this distribution have $\binom{n}{k}$?
- The average count for this distribution is $n\pi$
- The variance for this distribution is $n\pi(1 - \pi)$

```
set.seed(12324)
# Simulate 10 Binomial random variables with parameter n=2 and pi=0.5
rbinom(10, 2, 0.5)


##  [1] 1 2 2 1 2 0 0 1 1 1
```

Poisson Distribution

- The Poisson distribution is used to model the count of the occurence of events
- Classical application: Model for earthquakes
- The PMF is
$$P(K = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$
where $k = 0, 1, 2, \ldots$
- λ is the average number of events for this distribution
- λ is also the variance of this distribution

```r
set.seed(13224)
# Simulate 10 Poisson variates with m
rpois(10, 0.1)


##  [1] 0 1 0 0 0 0 1 0 0 0
```

Relationship between Binomial and Poisson Distribution

- Consider tossing the coin a large number of times

```
n = 1e+06
p = 1/n
```

- Note that we have $n = 10^6$ trials with a low success probability of $p = 10^{-6}$
- The expected number of events among these 10^6 trials is $n \times p = 1$. Why?
- Now simulate 99999 numbers from this binomial distribution

```
set.seed(9988)
x <- rbinom(B9, n, p)
length(x)

## [1] 99999
```

- What is the expected number of events (i.e., the expected number of events (among n trials) across $B = 99999$ simulations)?

```
mean(x)

## [1] 1.00055
```

Relationship between Binomial and Poisson Distribution

- Now compare the empirical distributions to the Poisson distributions

```
round(dpois(0:7, lambda = 1), 3)

## [1] 0.368 0.368 0.184 0.061 0.015 0.003 0.001 0.000

round(table(x)/B9, 3)

## x
##     0     1     2     3     4     5     6     7
## 0.367 0.369 0.183 0.061 0.016 0.003 0.000 0.000
```

Negative Binomial Distribution

- How many times do you have to flip a coin to get $r > 0$ events
- Model the number of *random* trials needed to get r events
- This distribution is called the negative binomial distribution
- The probability distribution is

$$P[K = k] = \binom{k + r - 1}{r - 1} \pi^r (1 - \pi)^k,$$

where $k = r, r + 1, r + 2, \ldots$

```
set.seed(13224)
# Simulate the number of trials needed to get k=5 events
rnbinom(10, 5, 0.1)


##  [1] 63 60 56 30 64 62 36 36 44 37
```

Mean and Variance of Negative Binomial

- A negative binomial distribution can be parameterized in terms of
  - r and p
  - or μ and σ^2
  - or μ and a dispersion parameter α (more on this later)
- The relationship between these two parametrizations is given by

$$\mu = r\frac{1-p}{p} \text{ and } \sigma^2 = r\frac{1-p}{p^2},$$

and

$$p = \frac{\mu}{\sigma^2} \text{ and } r = \frac{\mu^2}{\sigma^2 - \mu}$$

- If you provide r and p, you can calculate μ and σ^2
- Or, if you provide μ and σ^2, you can recover r and p.

Negative Binomial PMF in terms of μ and α

- The NB PMF parametrized in terms of p and r (the number of events) is

$$P[K=k] = \binom{k+r-1}{r-1}\pi^r(1-\pi)^k,$$

  where $k = r, r+1, r+2, \ldots$

- The NB PMF parametrized in terms of the mean μ and the dispersion parameter α is

$$P[K=k] = \frac{\Gamma[k+\alpha^{-1}]}{\Gamma[\alpha^{-1}]\Gamma[k+1]}\left(\frac{1}{1+\mu\alpha}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^k,$$

  where $k = 0, 1, \ldots$

- The variance is $\mu(1+\alpha\mu)$
- As α shrinks to 0 (no-dispersion), the distribution becomes Poisson

Negative Binomial PMF for RNA-Seq

- We will use the mean/dispersion parameter representation for RNA-Seq

  $$P[K=k] = \frac{\Gamma[k+\alpha^{-1}]}{\Gamma[\alpha^{-1}]\Gamma[k+1]} \left(\frac{1}{1+\mu\alpha}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{k},$$

  where $k = 0, 1, \ldots$
- The variance is $\mu(1+\alpha\mu)$
- IMPORTANT:
  - If $\alpha > 0$, then the variance is greater than the mean. Why?
  - As α shrinks to 0 (no-dispersion), the distribution becomes Poisson
- More on over-dispersion later

Means and Variances

Distribution	Support	Mean	Variance
Bernoulli(π)	0,1	π	$\pi(1-\pi)$
Binomial(n, π)	$0, 1, \ldots, n$	$n\pi$	$n\pi(1-\pi)$
Poisson(λ)	$0, 1, 2, \ldots,$	λ	λ
NB(p, r)	$r, r+1, r+2, \ldots,$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$
NB(μ, α)	$0, 1, \ldots,$	μ	$\mu(1+\alpha\mu)$

Multinomial Model

- Suppose that there are 3 urns
- n balls are to be randomly distributed among these M urns
- Let K_j denote the number of balls assigned to urns $j = 1, 2$ or 3
- Let $\pi_j = \mathbb{P}[X_i = j]$ denote the probability that ball $i = 1, \ldots, n$ is assigned to urn $i = 1, 2, 3$
- Finally let K_j denote the number of balls, among n, assigned to urn j
- (K_1, K_2, K_3) is said to have multinomial (trinomial) distribution with parameter $(3, \pi_1, \pi_2, \pi_3)$

Multinomial Model: A Check List

- Note that $K_1 + K_2 + K_3 = n$
- Why?
- Note that $\pi_1 + \pi_2 + \pi_3 = 1$
- Why?
- The support of K_j, the number of balls assigned to urn j, is $\{0, 1, \ldots, n\}$
- Why?
- The support if X_i, the urn to which ball i is assigned, is $\{1, 2, 3\}$
- Why?

Multinomial Model: Properties

- The PMF of X_i is

$$\pi_1 = \mathbb{P}[X_i = 1], \pi_2 = \mathbb{P}[X_i = 2] \text{ and } \pi_3 = \mathbb{P}[X_i = 3]$$

- The PMF of

$$\mathbb{P}[K_1 = k_1, K_2 = k_2, K_3 = k_3] = \frac{n!}{k_1!k_2!k_3!}\pi_1^{k_1} \times \pi_2^{k_2} \times \pi_3^{k_3}$$

- The PMF of K_1 is binomial with parameter (n, π_1)
- The PMF of K_2 is binomial with parameter (n, π_2)
- The PMF of K_3 is binomial with parameter (n, π_3)

Multinomial: Relationship to RNA-Seq

- Suppose that the genome has only three genes (the urns)
- n sequencing reads are to be mapped to these three genes
- K_1 is the number of reads mapped to gene 1
- K_2 is the number of reads mapped to gene 2
- $K_3 = n - K_1 - K_3$ is the number of reads mapped to gene 3
- The multinomial model provides a framework for thinking about the count model for (K_1, K_2, K_3)
- One can easily extend the multinomial distribution to arbitratry number of urns (genes)
- Caveat: Marginal PMFs do not account for overdispersion

Multinomial Simulation

```r
set.seed(3213)
### Number of balls
n <- 100
### The three urn probabilities
pik <- c(2, 1, 4)/7
pik
```

```
## [1] 0.2857143 0.1428571 0.5714286
```

```r
### Simulate two replicates from trinomial distribution with parameter
### (100,2/7,1/7,4/7)
K <- rmultinom(2, 100, pik)
K
```

```
##      [,1] [,2]
## [1,]   32   29
## [2,]   17   11
## [3,]   51   60
```

```r
### Add the two columns to verify they add up to n=100
apply(K, 2, sum)
```

```
## [1] 100 100
```