Simulació de dades RNAseq

La distribució binomial negativa

La distribució de Poisson suposa que la mitjana i la variància són iguals. Aquesta és una suposició molt forta. Una distribució de comptatge que permet diferir la mitjana i la variància és la distribució binomial negativa. Aprendre sobre la distribució binomial negativa ens permet generar i modelar tipus de comptatges més generals.

La variància de la distribució binomial negativa és una funció de la seva mitjana i un paràmetre de dispersió, k

$$\text{var}(Y) = \mu + \mu^2/k$$

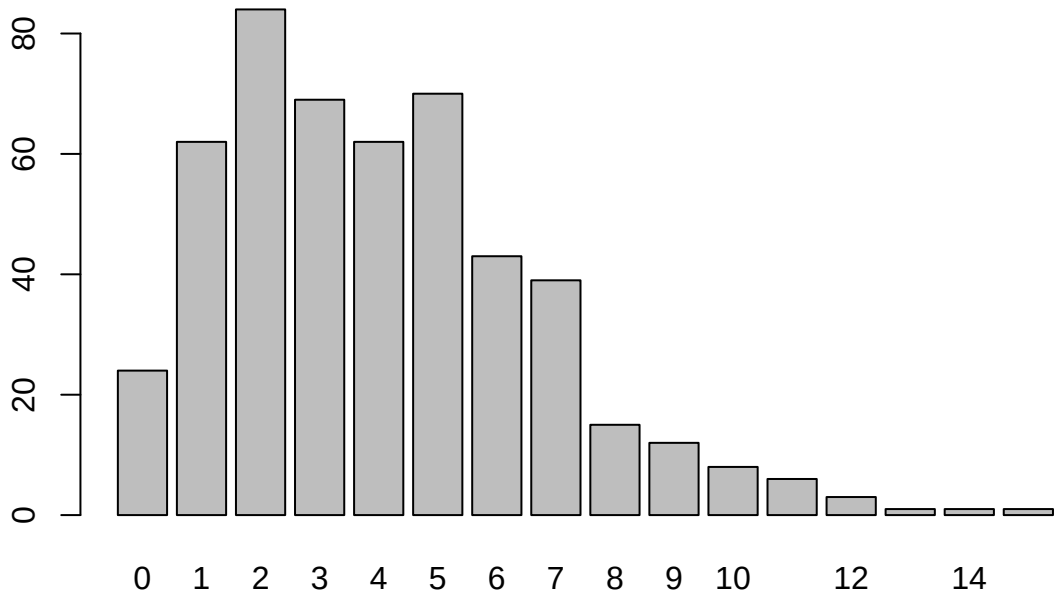
De vegades, k el trobareu indicat per una θ. A mesura que k es fa gran, la segona part de l'eqüació s'aproxima a 0 i convergeix a una distribució de Poisson.

Podem generar dades a partir d'una distribució binomial negativa mitjançant la funció `rnbinom`. El paràmetre de dispersió, k, s'especifica amb l'argument de mida. A continuació, generem 500 valors a partir d'una distribució binomial negativa amb $\mu = 4$ i $k = 5$:

```
y <- rnbinom(n = 500, mu = 4, size = 5)
table(y)
```

```
## y
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 16
## 24 62 84 69 62 70 43 39 15 12  8  6  3  1  1  1
```

```
barplot(table(y))
```




```
mu<-mean(y)
mu
```

```
## [1] 4.038
```

```
v<-var(y)
v
```

```
## [1] 7.170898
```

```
4+4^2/5
```

```
## [1] 7.2
```

Veiem que la variància és molt més gran que la mitjana. Per això, sovint diem que la distribució presenta una sobredispersió.

Una vegada més, generem un model senzill que produeix diferents recomptes basats en cada condició experimental en particular. Aquí teniu una manera d'aconseguir-ho fent servir el mateix model que abans, però aquesta vegada amb un paràmetre de dispersió que hem establert a 0,9. (Com que el paràmetre de dispersió es troba al denominador, els valors més petits de fet condueixen a més dispersió).

```
set.seed(5)
n <- 500
trt<- sample(c(0,1), size = n, replace = TRUE)
y_sim <- rnbinom(n = n,
                 mu = exp(-2 + 0.9 * (trt == 1)),
                 size = 0.05)
```

```
table(y_sim, trt)
```

```
##      trt
## y_sim   0   1
##    0  236 219
##    1   13   8
##    2    4   5
##    3    2   3
##    4    1   2
##    5    1   0
##    6    0   1
##    7    0   4
##    13   0   1
```

Regressió negativa binomial

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.2
```

```
m2 <- glm.nb(y_sim ~ trt)
summary(m2)
```

```
##
## Call:
## glm.nb(formula = y_sim ~ trt, init.theta = 0.06329572924, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4834  -0.4834  -0.3844  -0.3844   1.9754
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9656     0.2988  -6.579 4.73e-11 ***
## trt           0.8792     0.4080   2.155   0.0312 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.0633) family taken to be 1)
##
##     Null deviance: 129.13  on 499  degrees of freedom
## Residual deviance: 124.50  on 498  degrees of freedom
## AIC: 458.6
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.0633
##           Std. Err.:  0.0138
##
##  2 x log-likelihood:  -452.6020
```

El model sens dubte sembla "significatiu". Els coeficients estimats no estan massa lluny dels valors "veritables" de -2 i 0.9.

```r
set.seed(5)
n <- 5000
trt<- sample(c(0,1), size = n, replace = TRUE)
y_sim <- rnbinom(n = n,
                 mu = (2 + 0.9 * (trt == 1)),
                 size = 0.05)
```

```r
table(y_sim, trt)
```

```
##      trt
## y_sim    0    1
##   0   2144 1981
##   1     95  103
##   2     50   59
##   3     27   31
##   4     20   24
##   5     26   16
##   6     21   14
##   7     21   16
##   8     15   11
##   9      6   11
##   10     9   15
##   11     9    7
##   12     5   10
##   13     7    8
##   14     4    4
##   15     4    5
##   16     3    8
##   17     4   12
##   18     5    3
##   19     2    5
##   20     5    2
##   21     4    7
##   22     4    6
##   23     2    6
##   24     1    1
##   25     2    1
##   26     5    2
##   27     2    3
##   28     4    2
##   29     0    2
##   30     4    1
##   31     0    3
##   32     2    4
##   33     3    2
##   34     0    2
##   35     3    1
##   36     1    3
##   37     3    3
##   38     1    0
##   39     2    1
##   41     1    4
##   42     4    0
##   43     0    1
```

```
##   44     1    1
##   46     1    2
##   47     0    1
##   48     3    0
##   50     0    3
##   51     0    1
##   52     0    1
##   53     1    0
##   54     0    1
##   55     0    2
##   56     3    1
##   57     1    2
##   58     1    1
##   59     0    1
##   60     0    2
##   61     2    0
##   62     2    2
##   64     0    2
##   66     0    3
##   67     0    1
##   68     0    1
##   69     0    2
##   71     0    1
##   72     0    2
##   73     1    0
##   75     1    0
##   81     1    0
##   82     2    0
##   83     0    1
##   86     0    1
##   87     0    1
##   88     0    1
##   89     0    1
##   95     1    1
##   96     0    1
##   99     0    1
##   100    0    1
##   114    0    1
##   120    0    1
##   126    0    1
##   131    0    1
##   137    0    1
##   176    1    0
##   201    0    1
```

Regressió negativa binomial

```
library(MASS)
m2 <- glm.nb(y_sim ~ trt, link = identity)
summary(m2)
```

```
##
## Call:
## glm.nb(formula = y_sim ~ trt, link = identity, init.theta = 0.05005071743)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6396  -0.6396  -0.6023  -0.6023   2.9936
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8264     0.1638  11.150  < 2e-16 ***
## trt           1.1029     0.3132   3.522 0.000428 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.0501) family taken to be 1)
##
##     Null deviance: 1935.8  on 4999  degrees of freedom
## Residual deviance: 1922.2  on 4998  degrees of freedom
## AIC: 10472
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.05005
##           Std. Err.:  0.00199
##
##  2 x log-likelihood:  -10466.14900
```

El model sens dubte sembla "significatiu". Els coeficients estimats no estan massa lluny dels valors "veritables" de -2 i 0.9.