Mean-variance modeling

Kasper D. Hansen

Fall 2022

Counting and Poisson distributions

The most basic statistical distribution representing “counting” is the Poisson distribution.

Marioni (2008 Genome Res) and Bullard (2010 BMC Bioinformatics) established that if you sequence the exact same RNA-seq library multiple times, you get Poisson variation between the sequencing runs.

This result also holds for other type of sequencing (DNA, ChIP etc). This implies that the technical variation introduced by *sequencing* (excluding sample handling, library preparation etc) can be modeled as Poisson.

It is critical that it is *the exact same library*. This result does not mean that RNA-seq data is Poisson distributed.

Facts about the Poisson distribution

The standard model for sequencing is therefore

$$Y \sim \text{Poisson}(L\lambda)$$

where L is a known library size and λ is the unknown “expression” level of whatever you’re counting. L is used to be able to compare between runs with different sequencing depth.

We have

$$E(Y) = L\lambda, \quad \text{Var}(Y) = L\lambda, \quad \text{sd}(Y) = \sqrt{L\lambda}$$

In other words, the mean is the same as the variance. We have a mean-variance relationship.

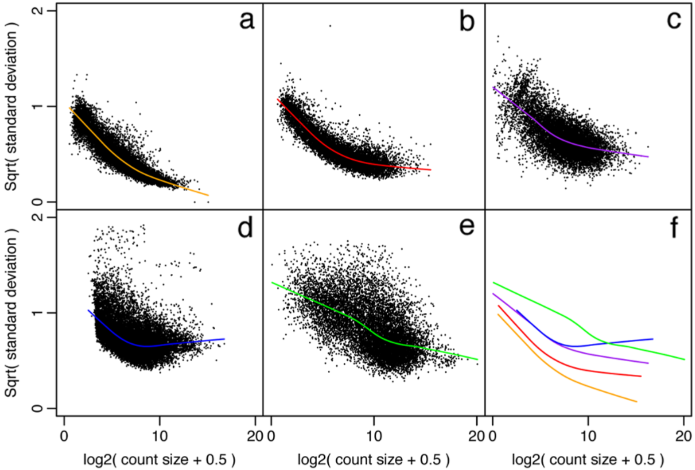
Does the variance increase with the mean?

If the mean is equal to the variance, it seems as if the variance gets bigger with the mean.

The result is different for log-transformed counts. Here, the variance will *decrease* as the mean increases.

For log-transformed data, we would expect a decrease (roughly following $1/x$) stabilizing at a level reflecting biological variance (details left out).

Mean-variance in RNA-seq data



a: SEQC
b: BL6 mice
c: simulation
d: LCLs
e: fruit fly development

Models

There are 2 roads to modeling mean-variance relationship.

- Use models built on the negative binomial model (DESeq2, edgeR)
- Use weighted linear models (limma-voom)

Experiments show that they are very similar in performance.