Overdispersion, Experimental Design in RNA-Seq

Joseph Abraham

Lecture VII RBP5793

Outline of Lecture

- Overdispersion
- Experimental Design in RNA-Seq

Useful References:

Statistical Design and Analysis of RNA Sequencing Data
P.L.Auer & R.W. Doerge, Genetics 185: 405-416 (2010)

Statistical Analysis of Next Generation Sequencing Data
S. Datta & D. Nettleton Eds Springer,
ISBN 978-3-319-07211-1

Overdispersion I

Our knowledge till now, the Poisson distribution describes count data, is defined through one parameter. This parameter λ is the value of the mean, and also of the variance. (Special Property of the Poisson Distribution). This is also a (strong) restriction on what kind of data can be (correctly) described by the Poisson Distribution.

Overdispersion II

Since the means and the variance are equal

we expect that if we take a sample from a

Poisson Distribution obtain the mean and variance

the value of

$$\frac{SampleVariance}{SampleMean}$$

should be equal to ?

Overdispersion III

Imagine we have the ratio of (variance/mean) for many

for many different samples below.

Which value seems to be an outlier ?

0,988	0,940	1,005	0,796	0,911
0,819	0,854	0,990	0,898	0,764
I,206	0,854	0,848	1,089	2,0182

Overdispersion III

Imagine we have the ratio of (variance/mean) for many

for many different samples below.

Which value seems to be an outlier ?

0,988	0,940	1,005	0,796	0,911
0,819	0,854	0,990	0,898	0,764
I,206	0,854	0,848	1,089	2,0182

All values (except the last) were obtained from a sample

of a single poisson distribution. The last was obtained by

joining the values from **2** Poisson distributions

with **different** means ! (Sample Heterogeneity !)

Overdispersion IV

In the example of patients receiving treatment for dengue during one calendar year will be a strong seasonal variation in the typical number of patients requesting treatment. There are other factors which cause a strong variation during the year in the average number of patients seeking treatment. Better to think of a typical number for certain months and a **different** value for other months. Values are from Poisson Distributions with **different** means.

Overdispersion V

For heterogeneous data from different Poisson distributions the value of the overdispersion given by

$$\frac{SampleVariance}{SampleMean}$$

will be larger than one. Or, if we find overdispersion we may have heterogeneity present. In RNA-Seq experiments this may be due to heterogeneity in the biological samples, even in the same group. Do we see overdispersion !

Overdispersion VI

Yes we do see overdispersion !

S. Anders & W. Huber Genome Biol. 2010; 11(10): R106.

Differential expression analysis for sequence count data

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218662/

Overdispersion VII

How to solve this problem ?

Anders & Huber Genome Biology 2010

. . . method based on the negative binomial distribution . . . implementation, DESeq, as an R/Bioconductor package . . .

Negative Binomial is also defined over count data (0,1,2 . . .)

Negative Binomial distribution has some similarity with

a combination of Multiple Poisson Distributions.

For a negative binomial distribution Variance $\geq$ Mean.

Overdispersion is included from the beginning !

Overdispersion VIII

Question: Imagine we want to study the number of traffic accidents per day on a certain street. We want the average, variance .... Is our data count data ?

Overdispersion VIII

Question: Imagine we want to study the number of traffic accidents per day on a certain street. We want the average, variance .... Is our data count data ? Suppose we obtain the data for 4 consecutive weeks. Suppose we try to use a Poisson Distribution to model our data. Can we expect overdispersion ? If yes why, if not why not ?

Experimental Design in RNA-Seq I

Auer & Doerge Genetics 185: 405-416 (2010)

https://www.ncbi.nlm.nih.gov/pubmed/20439781

in the absence of a proper design, it is essentially impossible to partition biological variation from technical variation. When these two sources of variation are confounded, there is no way of knowing which source is driving the observed results. No amount of statistical sophistication can separate confounded factors after data have been collected.

What is confounding in general ?

Experimental Design in RNA-Seq II

We assume that there are flow cells with multiple lanes

land in each lane sequencing is done separately.

Experimental Design in RNA-Seq II

We assume that there are flow cells with multiple lanes

land in each lane sequencing is done separately.

Are results from lanes independent of each other ?

Experimental Design in RNA-Seq II

We assume that there are flow cells with multiple lanes

land in each lane sequencing is done separately.

Are results from lanes independent of each other ?

Can use bar coding to sequence different samples in

the same lane (multiplexing). Important for efficiency and

here we demonstrate how multiplexing can be used as a quality control feature that offers the flexibility to construct balanced and blocked designs for the purpose of testing differential expression. Blocking and randomization needed not only at level of sample selection but also RNA extraction.

Experimental Design in RNA-Seq III

Imagine we want to compare between 7 treatments

we have 21 experimental units. Each treatment receives

3 experimental units after random assignment.

Is this design a balanced design ?

Experimental Design in RNA-Seq III

Imagine we want to compare between 7 treatments

we have 21 experimental units. Each treatment receives

3 experimental units after random assignment.

Is this design a balanced design ?

Completely Randomized Design ?

Experimental Design in RNA-Seq III

Imagine we want to compare between 7 treatments

we have 21 experimental units. Each treatment receives

3 experimental units after random assignment.

Is this design a balanced design ?

Completely Randomized Design ?

We use 3 flow cells of an Illumina Genome Analyzer.

Each flow cell has eight lanes. (One lane is for control).

Should we put all the experimental units with one treatment

in the same flow cell ?

Experimental Design in RNA-Seq III

Imagine we want to compare between 7 treatments

we have 21 experimental units. Each treatment receives

3 experimental units after random assignment.

Is this design a balanced design ?

Completely Randomized Design ?

We use 3 flow cells of an Illumina Genome Analyzer.

Each flow cell has eight lanes. (One lane is for control).

Should we put all the experimental units with one treatment

in the same flow cell ? If not, what is better ?

Experimental Design in RNA-Seq IV

Imagine we use in each of the three flow cells all of the 7 treatments. Each lane has the same treatment & 1 sample.

	L1	L2	L3	L4	L5	L6	L7	L8
FC1	T11	T21	T31	T41	C	T51	T61	T71
FC2	T12	T22	T32	T42	C	T52	T62	T72
FC3	T13	T23	T33	T43	C	T53	T63	T73

(T34 means the fourth sample receiving treatment 3)

Is each flow cell a block ?

Experimental Design in RNA-Seq IV

Imagine we use in each of the three flow cells all of the 7 treatments. Each lane has the same treatment & 1 sample.

	L1	L2	L3	L4	L5	L6	L7	L8
FC1	T11	T21	T31	T41	C	T51	T61	T71
FC2	T12	T22	T32	T42	C	T52	T62	T72
FC3	T13	T23	T33	T43	C	T53	T63	T73

(T34 means the fourth sample receiving treatment 3)

Is each flow cell a block ? If yes, is this good ?

Is this a complete block ?

Experimental Design in RNA-Seq V

Problem with previous design in practise.

Should not use the same treatment in the

same lane in each flow cell. Cannot separate the effects

of treatment and lanes ! Confounding again.

. . . "batch effects" and "lane effects." Batch effects include any errors that occur after random fragmentation of the RNA until it is input to the flow cell . . . Lane effects include any errors that occur from the point at which the sample is input to the flow cell until data are output from the sequencing machine

Can imagine lanes are blocks, need to randomize treatment

across lanes !

Experimental Design in RNA-Seq VI

To minimize lane effects we could randomize across lanes

Leads to a different assignment of treatments to lanes.

	L1	L2	L3	L4	L5	L6	L7	L8
FC1	T11	T21	T31	T41	C	T51	T61	T71
FC2	T73	T13	T21	T33	C	T42	T51	T62
FC3	T52	T61	T72	T12	C	T23	T31	T43

In previous lecture, blocking defined as homogeneity

between different biological replicates before observation.

Now blocking arises only after observation ! Whenever it arises,

must anticipate to maximize effects of treatment differences.

Experimental Design in RNA-Seq VI

If we treat lanes as blocks, can use bar codes and multiplexing to have a complete block design. Reduces the effects of confounding between treatments and lanes.

https://www.ncbi.nlm.nih.gov/pubmed/20439781

We know there may be a limit on the number of bar-codes permitted in a single lane. Assume this is 5.

Can we still perform a Complete Block design with 6 different treatments ? If not, what kind of design ?

Experimental Design in RNA-Seq VII

Useful design to know is the *Latin Square*.

Imagine we are testing 4 plant varieties (treatments)

in a field in which there are 2 blocking factors

The yield is affected by acidity (varies north to south)

and humidity (varies east to west). Want a balanced block

design in which we block against two factors.

Experimental Design in RNA-Seq VIII

One example of a four by four Latin square

The varieties are V1,V2,V3 and V4

V1	V2	V3	V4
V4	V1	V3	V2
V2	V4	V1	V3
V3	V2	V4	V1

Why is this design blocked in two directions ?

Experimental Design in RNA-Seq IX

https://link.springer.com/book/10.1007/978-3-319-07212-8