More Experimental Design, Poisson Distribution

Joseph Abraham

Lecture VI RBP5793

Outline of Lecture

- More Experimental Design
- Poisson Distribution

Useful Reference:

A First Course in the Design and Analysis of Experiments
Gary Oehlert, University of Minnesota. Available at
http://users.stat.umn.edu/ gary/book/fcdae.pdf

Experimental Design VII

From previous lecture we saw the concepts of

treatment, experimental unit, and response variable.

Also mentioned, the concepts of randomization

and replication. Two types of replication

Biological and Technical. What is the difference ?

Experimental Design VII

From previous lecture we saw the concepts of

treatment, experimental unit, and response variable.

Also mentioned, the concepts of randomization

and replication. Two types of replication

Biological and Technical. What is the difference ?

Which is better and why ?

Experimental Design VIII

Biological Replication means we have more Independent Experimental Units assigned to the various treatments. In a balanced design, we have the same number of experimental units per treatment. If not, design is unbalanced. If we have just one factor (*eg*. growth hormone) unbalanced designs may not be a problem. If we have two (or more) factors (*eg*. growth hormone and diet) unbalanced designs are much more complicated to analyse.

Experimental Design IX

Imagine we have an animal model to test

the combined effects of growth hormone and diet

on the weight gain of mice after a six week period.

Two levels for each factor. Four treatments in total.

4 Mice are assigned **at random** to each treatment.

16 mice in total.

Experimental Design IX

Imagine we have an animal model to test

the combined effects of growth hormone and diet

on the weight gain of mice after a six week period.

Two levels for each factor. Four treatments in total.

4 Mice are assigned **at random** to each treatment.

16 mice in total. Is this design balanced ?

Experimental Design X

For operational and convenience reasons we put the mice into different cages. Should we put all mice with the same treatment in the same cage ?

Experimental Design X

For operational and convenience reasons we put the mice into different cages. Should we put all mice with the same treatment in the same cage ?

Better to use two mice per cage. Each pair of mice in the one cage has the same treatment. Why should this be ?

Experimental Design X

For operational and convenience reasons we

put the mice into different cages. Should we put all

mice with the same treatment in the same cage ?

Better to use two mice per cage. Each pair of mice in the

one cage has the same treatment. Why should this be ?

Cages per treatment ? Experimental units are ?

Experimental Design X

For operational and convenience reasons we

put the mice into different cages. Should we put all

mice with the same treatment in the same cage ?

Better to use two mice per cage. Each pair of mice in the

one cage has the same treatment. Why should this be ?

Cages per treatment ? Experimental units are ?

During the six period one mouse dies. Is the design balanced.

Experimental Design X

For operational and convenience reasons we

put the mice into different cages. Should we put all

mice with the same treatment in the same cage ?

Better to use two mice per cage. Each pair of mice in the

one cage has the same treatment. Why should this be ?

Cages per treatment ? Experimental units are ?

During the six period one mouse dies. Is the design balanced.

Must analyze like an unbalanced design !

Experimental Design XI

In the previous design with 2 mice per cage

what are the observational units ?

Experimental Design XI

In the previous design with 2 mice per cage

what are the observational units ? Suppose we

used this design in another study which requires taking

a muscle sample from each mouse. We have 16 mice.

Now imagine we take **2** samples from each mouse.

32 samples in total. Same as biological replication ?

If not, why not ?

Experimental Design XII

In previous example, not all samples were independent ! Two samples from the same mouse receiving the same treatment are not independent observations ! For true biological replication need to increase the number of *independent* experimental units for each treatment.

In this case no increase in the numner of independent experimental units, even though the total number of measurements has increased. Example of **pseudoreplication.**

Experimental Design XIII

To see why pseudo replication is dangerous:

Consider the same quantity measured twice but independently (imagine two different technicians).

Results are X and Y but with some error.

We use the mean $= (\frac{X+Y}{2})$ but since X and Y have some error the result is not certain. We decide to specify a confidence interval as well the result.

Experimental Design XIV

The confidence interval is related to variability.

Variance also related to variability. What is the variance ?

Variance is $(\frac{Var(X)+Var(Y)}{4}) + \frac{Cov(X,Y)}{2}$

If observations are independent then $Cov(X, Y) = 0$ and

Variance is $(\frac{Var(X)+Var(Y)}{4})$ only.

Experimental Design XV

Now imagine we have *dependent observations*.

Now $Cov(X, Y) > 0$. This changes the variance.

For Independent Observations

variance $= (\frac{Var(X)+Var(Y)}{4})$

For Dependent Observations

variance $= (\frac{Var(X)+Var(Y)}{4}) + \frac{Cov(X,Y)}{2}$

Experimental Design XV

Now imagine we have *dependent observations*.

Now $Cov(X, Y) > 0$. This changes the variance.

For Independent Observations

variance $= (\frac{Var(X)+Var(Y)}{4})$

For Dependent Observations

variance $= (\frac{Var(X)+Var(Y)}{4}) + \frac{Cov(X,Y)}{2}$

Which Variance is larger ?

Which Confidence Interval is larger ?

Experimental Design XVI

More independent observations leads to smaller confidence intervals. If observations are *dependent* reduction is not the same ! In example with two observations per mouse (pseudoreplication) observations between the same mouse are dependent. Not the same as independent biological replication ! May however be advantageous to have two samples per mouse, but have to modify the analysis.

Experimental Design XVII

After randomization and replication, there is a third important concept in experimental design, **blocking**.

Imagine we wish to compare the effect of 3 different types of of fungicide on leaf fungus growth in the tree species.

We find a forest with many trees of that species. We choose some these trees at random. Each choosen tree is assigned 1 fungicide at random. The fungicide is applied to randomly choosen leaves which are studied for fungus growth.

Experimental Design XVIII

What are the experimental units ?

Experimental Design XVIII

What are the experimental units ? Observational units ?

Experimental Design XVIII

What are the experimental units ? Observational units ?

Treatments ? What are possible problems with this design ?

Experimental Design XVIII

What are the experimental units ? Observational units ?

Treatments ? What are possible problems with this design ?

Maybe the growth of fungus depends on other factors, such as moisture, sunlight, *etc.* which vary a lot between trees and which are hard to measure. However, within a given tree these factors do not vary very much. How to solve ?

Experimental Design XVIII

What are the experimental units ? Observational units ? Treatments ? What are possible problems with this design ? Maybe the growth of fungus depends on other factors, such as moisture, sunlight, *etc.* which vary a lot between trees and which are hard to measure. However, within a given tree these factors do not vary very much. How to solve ? We assign each fungicide at random to *different* leaves in the same tree ! Why ?

Experimental Design XIX

In this new design, what are the treatments, experimental units and observational units ?

When we treat leaves we first identify subsets of experimental units which are similar with regards to external factors.

These subsets are *blocks*. Treatment comparisons within the same block are less affected by external factors.

Differences in treatment are more visible within blocks.

Same treatment difference without blocking would require larger samples.

Experimental Design XX

Randomization, Replication and Blocking very old concepts.

Discussed in R. A. Fisher; *The Design of Experiments* (1935)

But still very relevant today !

Replication, randomization, and blocking are essential components of any well planned and properly analyzed design. RNA-Seq designs and analyses are no exception

Statistical Design and Analysis of RNA Sequencing Data

Paul Auer & Rebecca Doerge Genetics **185** 405-416 (2010)

Experimental Design XXI

Even with randomization, replication and blocking we still have many options and choices in experimental design. Simplest choice is CRD ,Completely Randomized Design (DIC, Delineamento Inteiramente Casualizado). Here we have a fixed number of experimental units. Select sample sizes. Assign units at random. If we have 8 units and we decide we need 5 with treatment A and 3 with B how to implement a random assignment ? (Each unit must have probaiblity 5/8 to get treatment A and 3/8 to get B).

Experimental Design XXII

Can be generalized to more than 2 treatments and different numbers of units per treatment. Design can balanced (balanceados) or the same number of units per treatment or unbalanced (não balanceado).

When the treatment can be thought of as a combined effect of more than one factor (*eg.* diet and drug) we have a factorial design. Can be balanced or unbalanced.

Experimental Design XXIII

If the factor A has three levels and B has 2 then any design in which there is at least one unit assigned to each of the 6 possible combinations is a full factorial design. Can be balanced or unbalanced and can be extended. Can imagine to be single factor with 6 levels but not this analysis is not always useful.

Experimental Design XXIV

Lynch, S. M. & J. J. Strain (1990) Nutrition Research 10, 449-460

6 treatments (3 milk levels and 2 copper levels)

were studied for their effect on rat liver function.

How does the milk level affect the outcome ?

How does the copper level affect the outcome ?

Is there an **interaction** between the two ?

(New feature for factorial designs).

Experimental Design XXV

Under the null hypothesis H_0 there is no effect.

and all 6 treatment means are the same. Under H_A

we can consider the main effects

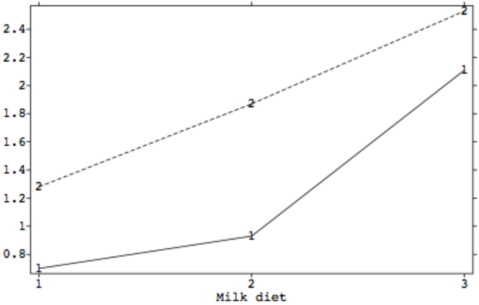
(milk and copper affect outcome independently)

or the interaction effect

(the effect of milk is different depending on the copper level).

Can visualize this on an interaction plot.

Experimental Design XXVI



Experimental Design XXVII

Lines not perfectly parallel. Difference in response
at different milk levels may depend on the copper treatment
Need to check the difference of differences by a
statistical test and find the confidence interval. This
analysis is much easier in the two factor picture. Easier
to see how H_0 may be untrue than with an analysis
of 6 different treatments.

Experimental Design XXVIII

The most common design used with blocking is the

Randomized Complete Block Design RCBD

(Delineameno em Blocos Completos Casualizados DBCC)

Each block contains at least one experimental unit

for each treatment. Within each block the assignment

of experimental units to treatments is random.

Block definition is not based on statistics ! Requires some

specialist knowledge of the problem.

Experimental Design XXIX

We have a cheese factory that works between 6:00 and 18:00 and we need to check that there is no difference in the quality of the cheese produced from 6:00 to 10:00, 10:00 to 14:00 and 14:00 h to 18:00 h. Fresh milk is supplied everyday.

We know that cheese quality is affected by milk quality.

Is it best to compare cheese produced between

6:00 to 10:00 h on Monday with that produced on Tuesday

between 14:00 and 18:00 ?

Experimental Design XXIX

We have a cheese factory that works between 6:00 and 18:00 and we need to check that there is no difference in the quality of the cheese produced from 6:00 to 10:00, 10:00 to 14:00 and 14:00 h to 18:00 h. Fresh milk is supplied everyday. We know that cheese quality is affected by milk quality. Is it best to compare cheese produced between 6:00 to 10:00 h on Monday with that produced on Tuesday between 14:00 and 18:00 ? How to define blocks ?

Experimental Design XXX

Sometimes cannot use all treatments under consideration in all blocks. If we have 3 types of eye drops (A,B,C) and we wish to compare their effects on human subjects, natural block is a person with only two eyes !

Blocks are Incomplete and cannot be completed using a different choice of block. For each block only one comparison possible. For making all possible comparisons can use the Balanced Incomplete Block Design.

If we have 4 participants with (A,B) 4 with (B,C) and 4 with (A,C) example of Balanced Incomplete Block Design.

Experimental Design XXXI

Sometimes we have two factors with multiple levels and enough experimental units. But random allocation is not possible, for logistical reasons. This leads to the Split Plot Experimental Design. Now have two factors involved, split plot and whole plot factors. Also two levels of experimental unit, split plot and whole plot.

Experimental Design XXXII

Imagine we have a field in which we wish to test

the effects of irrigation and plant type on yield.

We have two levels of irrigation and 3 plant types.

6 different treatments and 24 plots (experimental units).

However, it is difficult to assign adjacent plots to

different levels of irrigation (logistics). How to we proceed ?

Experimental Design XXXIII

Irrigation Levels I1 and I2, Plant types VA,VB and VC

For logistics must have same I level along each column

I1VA		I2VB		I2VC		I1VA
I1VC		I2VC		I2VA		I1VA
I1VB		I2VA		I2VA		I1VB
I1VA		I2VC		I2VB		I1VC
I1VB		I2VB		I2VB		I1VB
I1VC		I2VA		I2VC		I1VC

Example of a Split Plot Design.

Experimental Design XXXIII

Irrigation Levels I1 and I2, Plant types VA,VB and VC

For logistics must have same I level along each column

I1VA		I2VB		I2VC		I1VA
I1VC		I2VC		I2VA		I1VA
I1VB		I2VA		I2VA		I1VB
I1VA		I2VC		I2VB		I1VC
I1VB		I2VB		I2VB		I1VB
I1VC		I2VA		I2VC		I1VC

Example of a Split Plot Design. Full factorial ?

Experimental Design XXXIII

Irrigation Levels I1 and I2, Plant types VA,VB and VC

For logistics must have same I level along each column

I1VA		I2VB		I2VC		I1VA
I1VC		I2VC		I2VA		I1VA
I1VB		I2VA		I2VA		I1VB
I1VA		I2VC		I2VB		I1VC
I1VB		I2VB		I2VB		I1VB
I1VC		I2VA		I2VC		I1VC

Example of a Split Plot Design. Full factorial ? Balanced ?

Experimental Design XXXIII

Irrigation Levels I1 and I2, Plant types VA,VB and VC

For logistics must have same I level along each column

I1VA		I2VB		I2VC		I1VA
I1VC		I2VC		I2VA		I1VA
I1VB		I2VA		I2VA		I1VB
I1VA		I2VC		I2VB		I1VC
I1VB		I2VB		I2VB		I1VB
I1VC		I2VA		I2VC		I1VC

Example of a Split Plot Design. Full factorial ? Balanced ?

Randomize columns for I1 and I2 and then randomize

each column separately for VA, VB and VC.

Experimental Design XXXIV

In this design, Column is the Whole Plot Experimental Unit.

A row within a given column is the Split Plot Experimental Unit.

I is the whole plot factor and V is the split plot factor.

Experimental Design XXXIV

In this design, Column is the Whole Plot Experimental Unit.

A row within a given column is the Split Plot Experimental Unit.

I is the whole plot factor and V is the split plot factor.

Good design for comparing the effects of changing V while keeping I constant.

Experimental Design XXXIV

In this design, Column is the Whole Plot Experimental Unit.

A row within a given column is the Split Plot Experimental Unit.

I is the whole plot factor and V is the split plot factor.

Good design for comparing the effects of changing V while keeping I constant.

How to modify design to keep same whole plot and split plot experimental units but now use I as split plot factor and V as whole plot factor ?

Experimental Design XXXV

Same as before with a CRD

I1VA		I2VB		I1VA		I2VC
I1VC		I1VB		I2VB		I2VC
I2VC		I2VA		I1VA		I1VA
I1VA		I1VB		I2VB		I1VB
I2VB		I2VB		I1VC		I1VA
I2VC		I1VA		I1VC		I2VC

Randomly assign 24 plots, 4 each to 6 treatments

Full factorial ?

Experimental Design XXXV

Same as before with a CRD

I1VA		I2VB		I1VA		I2VC
I1VC		I1VB		I2VB		I2VC
I2VC		I2VA		I1VA		I1VA
I1VA		I1VB		I2VB		I1VB
I2VB		I2VB		I1VC		I1VA
I2VC		I1VA		I1VC		I2VC

Randomly assign 24 plots, 4 each to 6 treatments

Full factorial ? Balanced ?

Experimental Design XXXV

Same as before with a CRD

I1VA		I2VB		I1VA		I2VC
I1VC		I1VB		I2VB		I2VC
I2VC		I2VA		I1VA		I1VA
I1VA		I1VB		I2VB		I1VB
I2VB		I2VB		I1VC		I1VA
I2VC		I1VA		I1VC		I2VC

Randomly assign 24 plots, 4 each to 6 treatments

Full factorial ? Balanced ?

Poisson Distribution I

Upto now we have thought of gene expression in terms of the normal distribution. Observation does not have to be a whole number and the normal has mean and variance independent. The mean does not define the value of the variance !

Poisson Distribution I

Upto now we have thought of gene expression

in terms of the normal distribution. Observation does

not have to be a whole number and the normal

has mean and variance independent. The mean does

not define the value of the variance ! In RNA-Seq

the observations are counts and not continuous.

Need a statistical framework which is designed for counts.

Poisson Distribution II

Most straightforward distribution for describing counts

is the binomial distribution. Number of head is 4 or 7 . . .

but under no circumstances 3,6. Not Suitable for RNA-Seq !

Why ?

Poisson Distribution II

Most straightforward distribution for describing counts

is the binomial distribution. Number of head is 4 or 7 . . .

but under no circumstances 3,6. Not Suitable for RNA-Seq !

Why ? The observed value from a Binomial has

an upper limit. (Number of attempts). No such obvious

upper limit in RNA-Seq data. Need another distribution.

Poisson Distribution II

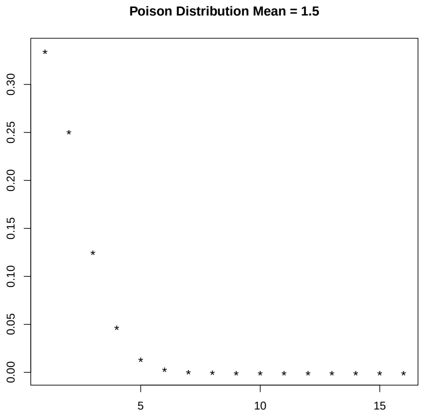
The Poisson Distribution describes count data and has no upper limit. Outcome can be 0,1,2,3 ... but only positive integers. Seems like a good starting point to describe RNA-Seq. Equation for the Poisson

$$\frac{e^{-\lambda}\lambda^{n}}{n!}$$

n is the observed value and takes values $0, 1, 2, 3 \ldots$

No upper bound on the value of n. λ is the average value.

Poisson Distribution III



Poisson Distribution IV

In previous slide we used $\lambda = 1, 2$

If we take one sample of size 200 from this

distribution what do we expect for the average value ?

Poisson Distribution IV

In previous slide we used $\lambda = 1, 2$

If we take one sample of size 200 from this

distribution what do we expect for the average value ?

If we take another sample (same size) from the same

distribution what do we expect for the average value ?

Poisson Distribution IV

In previous slide we used $\lambda = 1, 2$

If we take one sample of size 200 from this

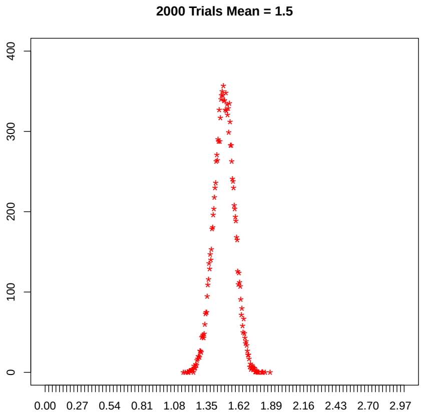
distribution what do we expect for the average value ?

If we take another sample (same size) from the same

distribution what do we expect for the average value ?

Can we expect a different value for a different sample ?

Poisson Distribution V


2000 Trials Mean = 1.5