

Introduction to Experimental Design

1/26/2009

Copyright © 2009 Dan Nettleton

1

Terminology

Experiment – An investigation in which the investigator applies some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

2

Terminology (continued)

Treatment – a condition or set of conditions applied to experimental units in an experiment.

Experimental Unit - the physical entity to which a treatment is randomly assigned and independently applied.

3

Terminology (continued)

Response Variable – a characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units.

Observational Unit - the unit on which a response variable is measured. There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

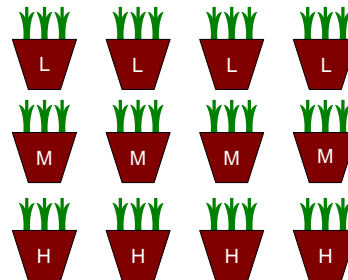
4

Example 1

- An experiment was conducted to study the effects of three soil moisture levels on gene expression in maize seedlings.
- A total of 36 seedlings were grown in 12 pots with 3 seedlings in each pot.
- The 3 soil moisture levels (low, medium, and high) were randomly assigned to the 12 pots with 4 pots for each soil moisture level.
- After three weeks, RNA was extracted from the above-ground tissues of each seedling.
- Each of the 36 RNA samples was hybridized to a microarray slide to measure gene expression.

5

A Cartoon Representation of the Experiment



6

Example 1 (continued)

1. Name the treatments in this experiment.
2. Name the experimental units in this experiment.
3. Name the observational units in this experiment.
4. Name the response variable or variables in this experiment.

7

Example 1 (continued)

1. Each of the three moisture levels represents a treatment.
2. The moisture levels were randomly assigned to the pots, so the pots are the experimental units. A pot consisting of 3 seedlings is one experimental unit.
3. Gene expression was measured for each seedling, so the seedlings are the observational units.
4. Each probe on the microarray slide provides one response variable. Thus, we will have several thousand response variables in this example.

8

Terminology (continued)

Response Variable – a characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units.

Explanatory Variable – a variable that can potentially be used to explain variation in a response variable.

9

Terminology (continued)

Factor – an explanatory variable that can take any one of two or more values.

Levels – the different values of a factor.

Treatment Factor – a factor whose levels are chosen and controlled by the researcher to understand how one or more response variables change in response to varying levels of the factor.

10

Terminology (continued)

Treatment Design – the collection of treatments used in an experiment.

Full Factorial Treatment Design – treatment design in which the treatments consist of all possible combinations involving one level from each of the treatment factors.

11

Example 2

- An experiment was conducted to gauge the effects of a drug and feed consumption on gene expression in rats.
- A total of 40 rats were housed in individual cages.
- Half of the 40 rats were randomly assigned to a calorie-restricted diet where daily feed rations contained approximately 50% of the calories normally consumed by rats of the type used in the experiment. The other 20 rats were provided with access to feeders that were always kept full so that their calorie intake was completely unrestricted.

12

Example 2 (continued)

- Within each diet group, four doses of an experimental drug (0, 10, 20, and 30 mg/kg body weight) were randomly assigned to rats with 5 rats per dose within each diet group.
- At the conclusion of the study, gene expression was measured for each rat using microarrays.

13

Example 2 (continued)

1. Name the treatment factors used in this experiment.
2. Name the levels of each factor.
3. Name the treatments used in this experiment.
4. Was a full factorial treatment design used?
5. Name the experimental units used in this experiment.
6. Name the observational units used in this experiment.

14

Example 2 (continued)

1. The treatment factors are *diet* and *drug*.
2. The levels of the factor *diet* are *restricted* and *unrestricted*. The levels of the factor *drug* are 0, 10, 20, and 30 mg/kg body weight
3. Each combination of diet and drug is one treatment. (R0, R10, R20, R30, U0, U10, U20, U30)
4. A full-factorial treatment design was used because all possible combinations of diet and drug were considered.
5. Each rat is an experimental unit and also an observational unit.

15

Terminology (continued)

Completely Randomized Design (CRD) – experimental design in which, for given number of experiment units per treatment, all possible assignments of treatments to experimental units are equally likely.

Block – a group of experimental units that, prior to treatment, are expected to be more like one another (with respect to one or more response variables) than experimental units in general.

Randomized Complete Block Design (RCBD) – experimental design in which separate and completely randomized treatment assignments are made for each of multiple blocks in such a way that all treatments have at least one experimental unit in each block.

16

Three Fundamental Experimental Design Concepts Attributed to R.A. Fisher

Randomization – random assignment of treatments to experimental units.

Blocking – grouping similar experimental units together and assigning different treatments within such groups of experimental units.

Replication – applying a treatment independently to two or more experimental units.

17

Example 3

- Suppose an experiment is to be conducted to study the effects of 5 treatments (A, B, C, D, and E) on gene expression in dairy cattle.
- A total of 25 GeneChips and a total of 25 cows, located on 5 farms with 5 cows on each farm, are available for the experiment.
- Which of the following designs is better from a statistical standpoint?

18

Example 3 (continued)

- Design 1: To reduce variability within treatment groups, randomly assign the 5 treatments to the 5 farms so that all 5 cows on any one farm receive the same treatment. Measure gene expression using one GeneChip for each cow.
- Design 2: Randomly assign the 5 treatments to the 5 cows within each farm so that all 5 treatments are represented on each farm. Measure gene expression using one GeneChip for each cow.

19

Example 3 (continued)

Design 1	Design 2
Farm 1: B B B B B	Farm 1: A B E D C
Farm 2: D D D D D	Farm 2: E D A C B
Farm 3: A A A A A	Farm 3: C D E A B
Farm 4: E E E E E	Farm 4: A B E C D
Farm 5: C C C C C	Farm 5: C A D B E

20

Example 3 (continued)

1. Name the observational units in each design.
2. Name the experimental units in each design.
3. Is blocking used for either design? If so, describe the blocks.
4. Describe the level of replication for each experimental design.

21

Example 3 (continued)

1. Cows are the observational units in both designs.
2. Farms are the experimental units in Design 1, and cows are the experimental units in Design 2.
3. Design 2 is a randomized complete block design (RCBD) with a group of 5 cows on a farm serving as a block of experimental units.
4. Design 1 has no replication because there is only 1 experimental unit for each treatment. Design 2 has 5 replications per treatment.

22

Example 3 (continued)

- Design 2 is by far the better design.
- We can compare treatments directly among cows that share the same farm environment.
- With Design 1 it is impossible to separate differences in expression due to treatment effects from differences in expression that might be due to farm effects.

23

The Importance of Randomization

Consider a field experiment intended to compare the yield of two corn varieties (A and B).

Suppose the field is divided into 20 plots that run from one end of the field to the other.

Is there anything wrong with the following assignment of varieties to field plots?

A B A B A B A B A B A B A B A B A B A B A B A B

24

The Importance of Randomization

Suppose a researcher would like to compare the effects of three treatments on gene expression in mice.

The researcher has a large cage containing 18 female mice to use as experimental units.

A lab technician reaches into the cage and picks up one mouse at a time for placement in individual cages that that will house the mice during the experiment.

25

The Importance of Randomization

The first 6 mice picked up by the researcher are assigned to treatment A.

The next 6 are assigned to treatment B.

The last 6 are assigned to treatment C.

Do you see any problem with this strategy of assigning treatments to experimental units?

26

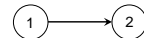
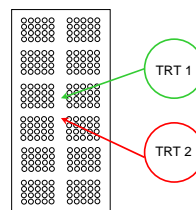
Generating Random Assignments in R

- For the mouse example (CRD),
`ra=cbind(1:18,sample(rep(c("A","B","C"),6)))`
`colnames(ra)=c("ID","TRT")`
`write.table(ra,"ra.csv",quote=F,sep=" ",row.names=F)`

- For the field experiment (RCBD),
`ra=matrix(rep(c("A","B"),10),nrow=10,byrow=T)`
`ra=apply(ra,1,sample)`
`ra=cbind(1:10,t(ra))`
`colnames(ra)=c("Block","Plot1","Plot2")`
`write.table(ra,"ra.csv",quote=F,sep=" ",row.names=F)`

27

Microarray Experimental Design Notation



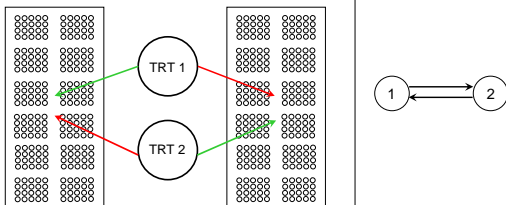
Each circle is an experimental unit labeled with its treatment.

Each arrow is a slide connecting experimental units hybridized to it.

The direction of the arrow denotes dye assignment.

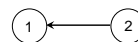
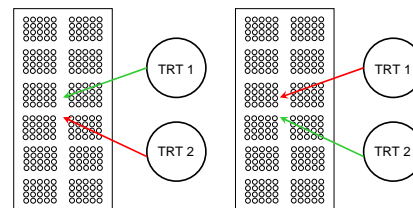
28

Microarray Experimental Design Notation



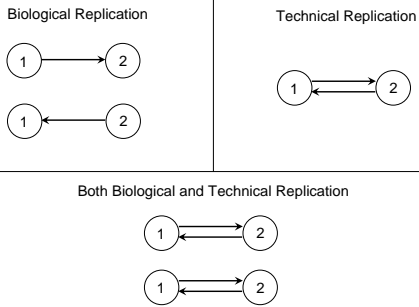
29

Microarray Experimental Design Notation



30

Biological Replicates vs. Technical Replicates



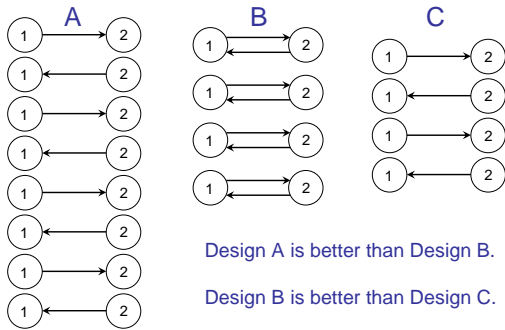
31

Some General Microarray Experimental Design Advice

- Use as much biological replication as is affordable.
- If the number of microarray slides or GeneChips is the limiting factor, measure each sample only once. Measuring any one sample more than once reduces the degree of biological replication that is possible, and this reduces the power to detect differential expression.
- If the number of biological replications is the limiting factor, measuring each experimental unit multiple times can improve precision, but this technical replication is no substitute for biological replication.

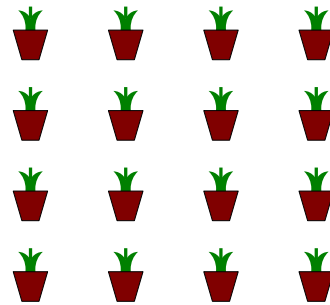
32

Biological Replication Should be Maximized



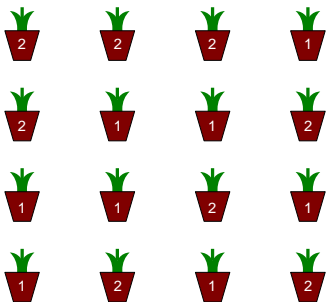
33

Example 4: Two-Treatment CRD



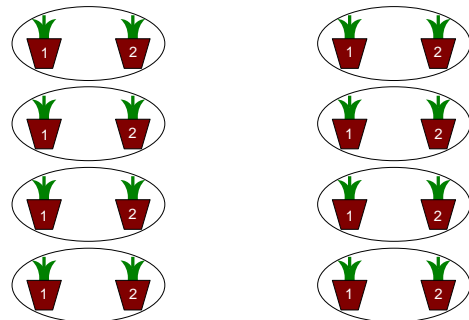
34

Assign 8 Plants to Each Treatment Completely at Random



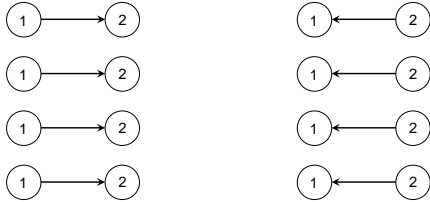
35

Randomly Pair Plants Receiving Different Treatments



36

Randomly Assign Pairs to Slides Balancing the Two Dye Configurations



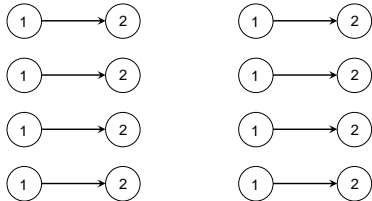
37

Confounding

- Confounding occurs when the effects of two or more explanatory variables (on a response variable of interest) cannot be distinguished from one another.
- Confounding can be problematic or useful depending on the situation.
- In Design 1 from Example 3, the effects of farms and treatments were completely confounded. This was very problematic.

38

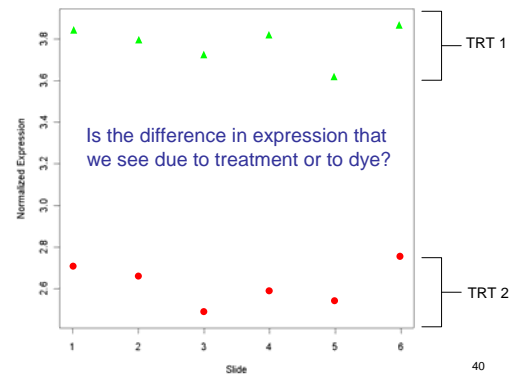
Another Example of Problematic Confounding



Treatment is completely confounded with dye in this two color microarray experiment.

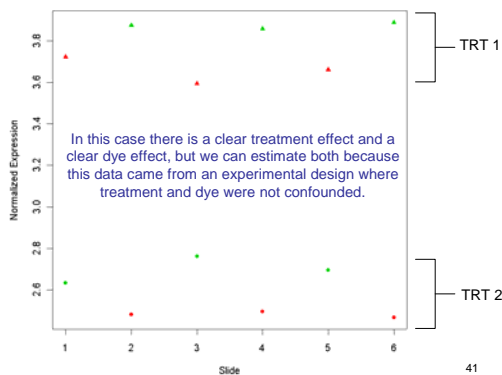
39

Suppose we see data as follows:



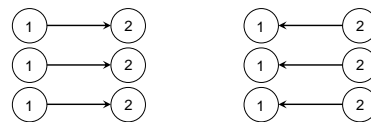
40

Actual Data from a Dye-Balanced CRD as in Example 4



41

An Example of Useful Confounding



The technician that executed this experiment hybridized one randomly selected slide on each of 6 days.

The sample assigned to the Cy3 dye was always processed before the Cy5 sample.

Process order is completely confounded with dye.

42

An Example of Useful Confounding (continued)

The dye effect that we will include in our model is actually dye confounded with process order.

This confounding is useful because by including one factor in our statistical model, we can simultaneously account for two nuisance factors that can affect expression measures.

This technique can only be used for factors whose effects are not of scientific interest.

If for some reason we want to separately estimate dye effects and order effects, then we would have to design and analyze the experiment differently.

43

An Example of Useful Confounding (continued)

The effects of day and slide are also completely confounded.

This is useful if we don't care about separately estimating slide-to-slide variation or day-to-day variation.

By including *slide* or, equivalently, *day* in the model used for statistical analysis, we will simultaneously account for both sources of variation.

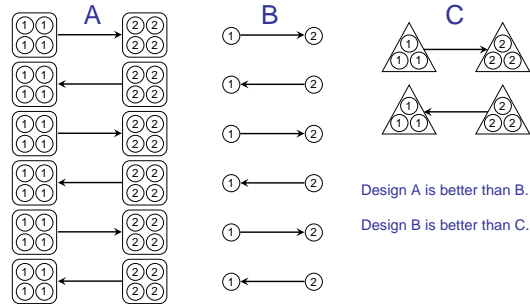
44

Pooling Tissues or RNA Samples

- Pooling of tissues or RNA samples is sometimes necessary to obtain sufficient RNA for hybridization.
- Even when pooling is not necessary, it can be beneficial because RNA levels are less variable across multiple pools than across multiple individual samples.
- When the number of slides is a limiting factor and experimental units are inexpensive, consider pooling to enhance power for detecting differentially expressed genes.

45

Designs A and C measure pools rather than individual samples.



46