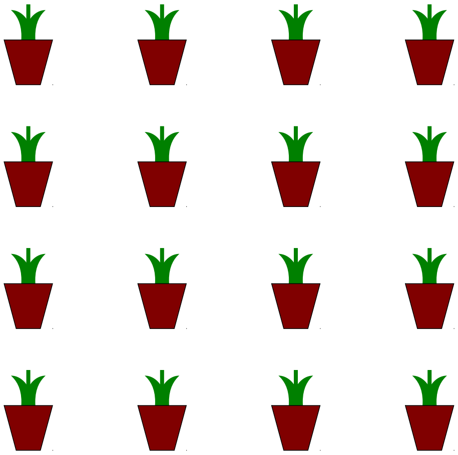
Kim-Anh Lê Cao April 2025

Introduction to experimental design: exercises

1 Soil moisture



Exercise 1: soil moisture experiment

An experiment was conducted to study the effects of three soil moisture levels on gene expression in maize seedlings (see Figure 1). 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.

1. Name the treatments in this experiment.
   The treatments in this experiment are the three Soil Moisture levels - Low, Medium, High. In statistical jargon, we would say that there is one *treatment factor*, namely Soil Moisture. The *levels* of this treatment factor are Low, Medium, High.

2. Name the experimental units in this experiment.
   The experimental units in this experiment are the pots (or equivalently, the groups of 3 seedlings contained *within* each pot).

   The individual seedlings in a pot are *not* the experimental units, because the treatments (Soil Moistures) are not being applied to the seedlings *independently* - if we know that one seedling in a particular pot has received the Low soil mositure treatment, then we immediately that the other seedlings in that pot have received the same treatment!

3. Name the observational units in this experiment.
   The description indicates that an RNA sample is obtained from each distinct seedling (of which there are 36), and these 36 are each hybridized to a microarray slide to measure gene expression. Thus, for a given RNA sequence, we will have 36 observations, one from each of the 36 seedlings. So, the *seedlings* are the observational units.

4. Name the response variable or variables in this experiment.
   Each probe on the microarray slide provides one response variable. Thus, we will have several thousands of response variables in this example.

   Discussion aside - the common issue that comes up with high-throughput experiments is when there is confusion over ’number of responses’ and ’number of replicates’. Each microarray may allow us to measure thousands of responses (expression of different distinct RNA sequences), but this does not mean that there is huge amounts of (biological) replication.

2 Plants

Establish a strategy to assign to 8 plants to either of two treatments completely at random.
The best way would be to use a random generator, online or using a software. Or for a two treatment design an unbiased coin. Just remember that a human cannot consciously generate random numbers!

3 Puppies

An investigator wants to examine the effectiveness of 2 drugs A and B for controlling heartworms in puppies. Veterinarians gave conjectures that the effectiveness of the drugs may depend on a puppy’s diet. Three different diets are combined with the two drugs. Also, the effectiveness of the drugs may depend on a transmitted inherent protection against heartworm obtained from the puppy’s mother. We consider 4 litters, with 6 puppies in each litter.

1. What are the factors in this experiment, how many treatments are compared?
   Diet is factor 1 and drug is factor 2, we have a 3×2 factorial treatment structure consisting of 6 treatments.

2. What is the blocking factor?
   Four litters of puppies consisting of 6 puppies each were selected to serve as a blocking factor in the experiment because all puppies within a given litter have the same mother.

3. Describe the design in a table.
   The six factor level combinations (treatments) were randomly assigned to the six puppies within each of the four litters. The design is a Randomized Complete Block Design in which the blocks are the litters and the treatments are the six factor-level combinations of the 3×2 factorial treatment structure.

	Litter			
Puppy	1	2	3	4
1	A-D1	A-D3	B-D3	B-D2
2	A-D3	B-D1	A-D2	A-D2
3	B-D1	A-D1	B-D2	A-D1
4	A-D2	B-D2	B-D1	B-D3
5	B-D3	B-D3	A-D1	A-D3
6	B-D2	A-D2	A-D3	B-D1

4 Dairy cattle

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 lanes (on one chip) and a total of 25 cows, located on 5 farms with 5 cows on each farm, are available for the experiment.

- 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 lane 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 lane for each cow.

For each design, answer the following questions

1. Represent the design in a table.

   (a) Design 1

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

(b) Design 2

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

2. Name the observational units in each design.
   Cows are the observational units in both designs.

3. Name the experimental units in each design.
   Farms are the experimental units in Design 1, and cows are the experimental units in Design 2.

4. Is blocking used for either design? If so, describe the blocks.
   Design 2 is a randomized complete block design with a group of 5 cows on a farm serving as a block of experimental units.

5. Describe the level of replication for each experimental design.
   Design 1 has no replication because there is only 1 experimental unit for each treatment (usually we assume that each block consists of very homogeneous observational units and the replications concerns the number of experimental units). Design 2 has 5 replications per treatment.

6. Which of the following designs is better from a statistical standpoint?
   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.