Section 2.2:

Sampling in Enumerative Studies

Recap

Data Collection

General Principles

Section 2.2: Data Collection in Enumerative Studies

- Enumerative studies: well defined population and sample taken from that population.
- Most useful way to create the sample: **Simple Random Sampling** - any group of *n* objects has the same chance of composing the sample as any other group of *n* objects.
- Suppose we have the alphabet (A, B, C, ..., Z) and wish to use simple random sampling to draw 3 letters. This means that the trio "F, M, Q" and the trio "A, B, C" have the same chance of being the letters that compose our sample.
- "Random" is tough to do correctly on your own. There are a few simple tools, like *random number tables* or *pseudo random number generators*, that help us.

Using Random Numbers to Get a Sample

- These tables are generated randomly - each place on the table is equally likely to be filled by any one of the numbers 0 - 9.
- The tables are created by taking advantage of some process that is physically random - radioactive decay or white noise for instance.
- RANDOM.org for example uses the amount of atmospheric static to generate the numbers.
- To use the randomly generated numbers to get a sample, simply assign a unique value to each item and take the items as they are generated.

Using a Random Number Table

For a simple random sample of size (n) from a population of size (N),

1. let m be the length in digits of N (for instance, if $N = 1032$ then $m = 4$
2. assign each item in the population a value between 1 and N
3. starting on the top left, box the first m digits. If the value is between 1 and N then take the item with that value assigned to it as part of your sample. Otherwise, box the next four letters.
4. continue until you have selected n items

m= 4

Table 2.2

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379
59069	01722	53338	41942	65118	71236	01932	70343	25812	62275
54107	58081	82470	59407	13475	95872	16268	78436	39251	64247
99681	81295	06315	28212	45029	57701	96327	85436	33614	29070

Recap

Data Collection

General Principles

Get a SRS

Ex: SRS tools

Using a Random Number Table

Take a simple random sample of size 3 from a set of 25 microprocessors using Table 2.2:

1. In this case $m = 2$, and we are given $n = 3$ and $N = 25$.
2. Each microprocessor gets given a number from 1 to 25.
3. Begin selecting the items

Table 2.2

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379
59069	01722	53338	41942	65118	71236	01932	70343	25812	62275
54107	58081	82470	59407	13475	95872	16268	78436	39251	64247
99681	81295	06315	28212	45029	57701	96327	85436	33614	29070

4. **Result**: select the microprocessors labeled 12, 15, and 05

Population: $x_1, x_2, \ldots, x_{N=25}$

sample: x_{12}, x_{15}, x_5
$n = 3$

Recap

Data Collection

General Principles

Get a SRS

Ex: SRS tools

Using pseudo-random numbers

```
sample(1:25,3) # some R code to get SRS of size 3
```

R output

sample(LETTERS,3)

[1] "S" "D" "Y"

sample(letters, 3)

[1] "q" "m" "i"

sample(1:25, 3)

[1] 20 19 24

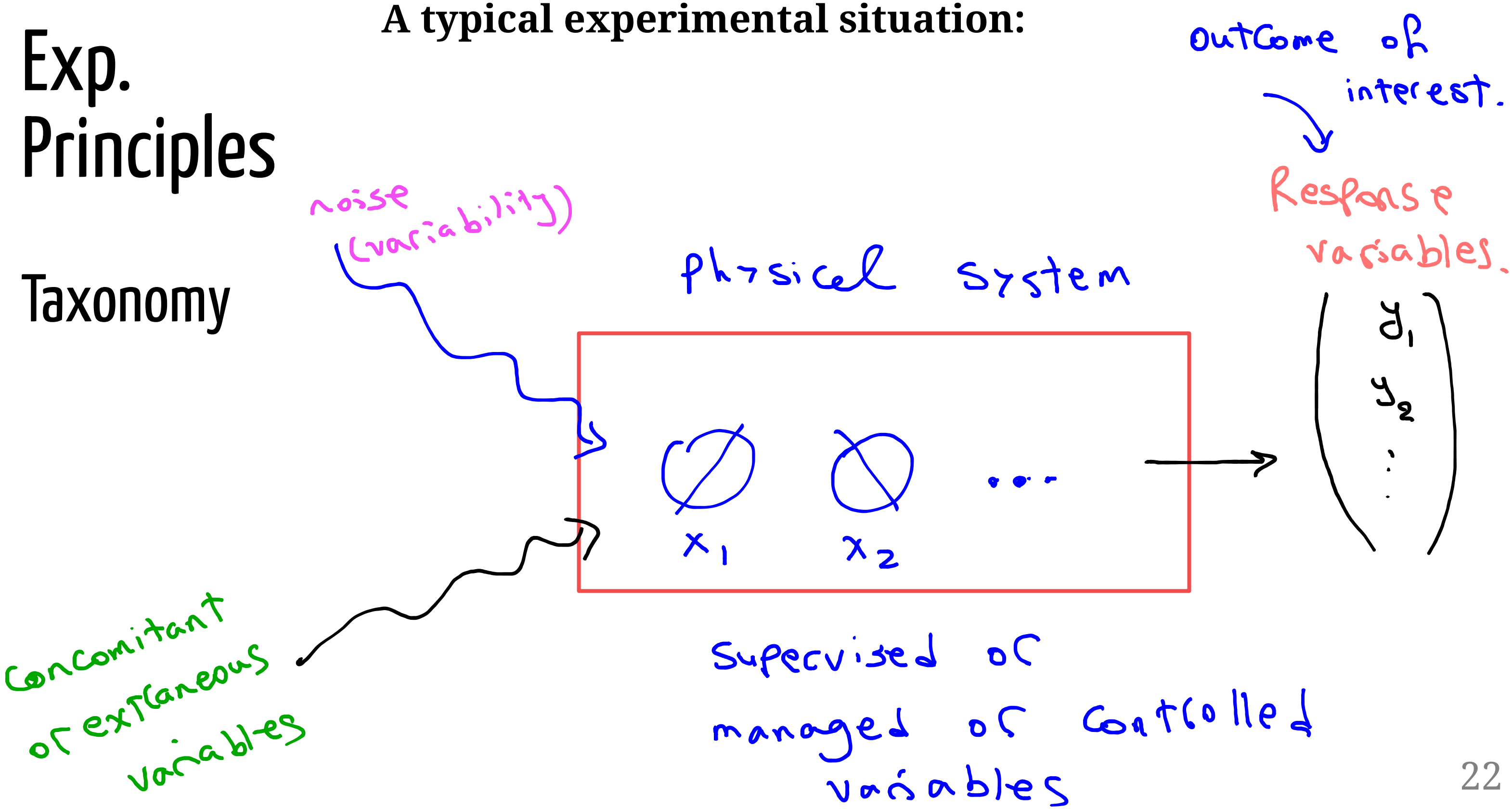
Section 2.3

Principles for Effective Experimentation

Effective experimentation

Purposefully changing a system and observing what happens as a result is a principled way of learning how a system works.

A typical experimental situation:



Recap

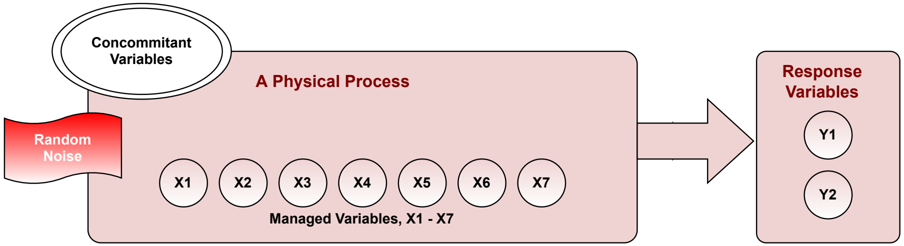
Data Collection

Exp. Principles

Taxonomy

Effective experimentation

Taxonomy of variables



Planning an experiment is complicated. There are typically many different characteristics of the system an engineer is interested in improving and many variables that might influence them. Some terminology is needed.

Response Variable

> A **response variable** in an experiment is one that is monitored as characterizing system performance/behavior.

Recap

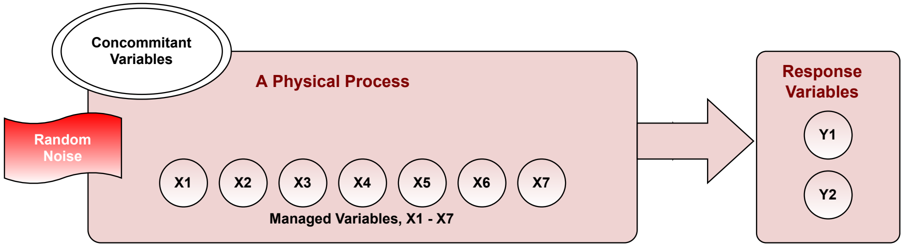
Data Collection

Exp. Principles

Taxonomy

Effective experimentation

Taxonomy of variables



Supervised variable

> A **supervised (or managed) variable** in an experiment is one over which an investigator exercises power, choosing a setting or settings for use in the study.
>
> When a supervised variable is held constant (has only one setting), it is called a *control variable*. When a supervised variable is given several settings in a study, it is called an *experimental variable*.

Recap

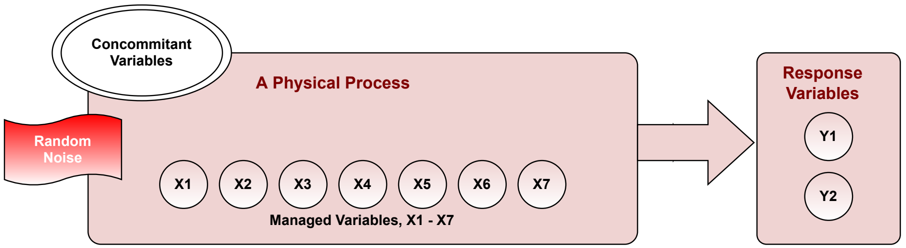
Data Collection

Exp. Principles

Taxonomy

Effective experimentation

Taxonomy of variables



concomitant variable

> A **concomitant (or accompanying) variable** in an experiment is one that is observed but is neither a primary response variable nor a managed variable.
>
> Such a variable can change in relation to either experimental or unobserved causes and may or may not itself have an impact on a response variable.

Recap

Data Collection

Exp. Principles

Taxonomy

Effective experimentation

Taxonomy of variables

Example:[Chemical purity]

Suppose you want to know about the effect of two different reactants (A and B) on the purity of a chemical for a given mixing speed and batch size. Reactant A has 2 levels (a_1 and a_2) and reactant B also has 2 levels (b_1 and $b2$).

What are the response variables, controlled variables, experimental variables, and concomitant variables?

Response : purity of chemical

Controlled variable(s) : speed & batch size

Experimental variable(s) : reactant A, B

Concomitant variable(s) : air temp.
pressure

Recap

Data Collection

Exp. Principles

Taxonomy

Example:[Wood joint strength, pg. 39]

Dimond and Dix experimented with three different woods and three different glues, investigating joint strength properties. Their primary interest was the effects of wood type and glue type on joint strength in a tension test and joint strength in a shear test. In addition, they found that the strengths were probably related to the variables drying time and pressure, so they hold these two variables constant. They also observed that variation in strengths could also have originated in properties of the particular specimens glued, such as moisture content although they haven't utilized this variable in the analysis of the data.

What is a full/complete factorial study for this experiment? What are the response variables, controlled variables, experimental variables, and concomitant variables?

response variable(s) : joint strength in tension test
" " " shear "

controlled variable(s) : drying time
pressure

Experimental variable(s): wood type (1, 2, 3)
glue type (A, B, C)

Concomitant variable(s): moisture content.

Full factorial study: $(\# \text{levels})^{\# \text{factors}}$

$= 3^2 = 9$

glue type

		g_A	g_B	g_C
wood type	w_1	$w_1 g_A$	$w_1 g_B$	$w_1 g_C$
	w_2	$w_2 g_A$	$w_2 g_B$	$w_2 g_C$
	w_3	$w_3 g_A$	$w_3 g_B$	$w_3 g_C$

Recap

Data Collection

Exp. Principles

Taxonomy

Extraneous variables

Extraneous variables are undesirable variables that influence the relationship between the variables that an experimenter is examining. Extraneous variables that vary with the levels of the independent variable are the most dangerous type in terms of challenging the validity of experimental results. These types of extraneous variables have a special name, **confounding variables**.

There are three basic ways to handle extraneous variable:

- Treat them as controlled variables (Hard to extend results)
- Handle them as experimental variables and create a second homogeneous environment to compare levels of priamary experimental variables (blocking)
- Randomization

Recap

Data Collection

Exp. Principles

Taxonomy

Blocking Variable

> A **block** of experimental units, experimental times of observation, experimental conditions, etc. is a homogeneous group within which different levels of primary experimental variables can be applied and compared in a relatively uniform environment.

Randomization

> **Randomization** is the use of a randomizing device at some point where experimental protocol is not already dictated by the specification of the supervised variables.
>
> Often it means that assigning experimental units to the experimental conditions at random.

Recap

Data Collection

Exp. Principles

Taxonomy

Example:[Heat treating gears, cont'd]

A process engineer is faced with the question, "How should gears be loaded into a continuous carburizing furnace in order to minimize distortion during heat treating?"

1

There are two types of methods for loading, laying or hanging the gears. The thrust face runout (0.0001 in) is a measure of distortion.

2

What are the response variables, controlled variables, experimental variables, and extraneous variables? How would you use handle the extraneous variable (three ways)?

- Response variable: Thrust face runout.
- Controlled variable: —
- Experimental variable: loading gears
  - hang
  - lay
- Extraneous variable: gear type

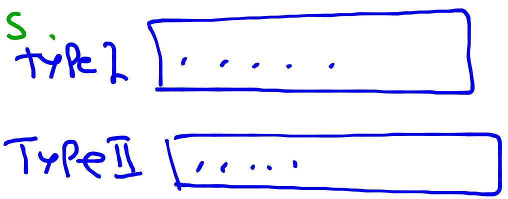
How to handle the extraneous variables?

① Controlled variable:

use only one type of gears, exactly the same for the whole study.

② Blocking:

If there are two (or more) types of gears, apply both hanging & laying methods to both types of gears.



③ Randomization:

If there are two types of gears, randomly assign each to each method.

Recap

Data Collection

Exp. Principles

Taxonomy

Some key issues of data collection

- **Comparative study:** Need a point of reference. e.g. studying the strength of a new alloy, may need the strength of the existing alloy.
- **Replication:** Need evidence that a study is repeatable or reproducible (resutls are not by a chance or mistake)
  under the same setting, need to collect the data more than once.
- **Allocation of resources:** Need to plan ahead
  Spend your resources (time/ money/ lab space/ materials) sequentially. i.e. extend your experiment gradually if possible to get priliminary results.

If data variablity is high, need more data.

Recap

Data Collection

Exp. Principles

Taxonomy

Extraneous Vars

> **Common Advice**: Block what you can control and randomize the rest (common, not necessarily good though, what can be controlled not universal).

Recap

Section 2.3: Principles of Effective Experimentation

Recap

Terminology

We described an ideal simple experiment and defined a few associated terms (2.3.1, 2.3.2)

- **Managed variable**: variables where we choose the value
    - **Controlled variable**: a managed variable that only takes one value throughout our experiment.
    - **Experimental variable**: a managed variable taking different values for different runs.
- **Response variable**: the output value; the variables whose values we wish to effect using our experimental variables
- **Concomitant variable**: Variables that we record but are not of interest.
- **Extraneous variable**: All other changes in our system.

Recap

Terminology

Comparative Studies

Comparitive Studies

We discussed the importance of **Comparative studies** (2.3.3)

- def: a study in which the goal is to compare two or more approaches/methods/ideas/etc.
- As always, there are many things that we can not control when we make measurements and collect data
- When we want to compare a new method to a old method, we need to be aware that the uncontrolled circumstances that existed when we first studied the old method do not exist anymore
- Take away: in order to *know* that difference in results are due to the difference in the method used (and not the difference in the uncontrolled circumstances) we must collect new data on both methods - this way, both methods be studied under the same uncontrolled circumstances.

Recap

Terms

Comparative Studies

Techniques

Techniques for Dealing with Extraneous Variables

Some aspects of the environment can not be controlled, even though they may effect the results we see. We call those aspects **extraneous variables**.

Though we can not *control* extraneous variables, we can plan our experiment to minimize their impact (2.3.2, 2.3.4)

Note: These techniques are part of how the experiment is *designed* - we decide this before any data is collected.

Technique 1: Blocking

- Designing the experiment with homogeneous mini-environments.
- In this way, regardless of the random/uncontrolled events that occur in the mini-environment, we know every observation experienced the same event.
- When collecting the data, we record the identity of the block used. Since the block used will have different values depending on the observational unit, we call the block identifier a "blocking variable"

Techniques for Dealing with Extraneous Variables

Technique 2: Randomization

- Using random assignment at all possible chances to "average out" the systematic changes that occur as we perform each run of our experiment.
- Used to choose the assignment of order, location, worker, partners, etc.

Techniques for Dealing with Extraneous Variables

Technique 2: Randomization, continued

Example: A chemist performs 20 runs of a synthesis, 10 using substrate A and 10 using substrate B. It is believed that the chemist could become more adept with each run.

Attempt 1: No randomization on order:

- **Plan**: the chemist performs all 10 syntheses using A and then performs all 10 syntheses using B.
- **Impact**: we have a changing extraneous variable ("adeptness") that will benefit the last runs
- **Result**: we can't tell the difference between whether our results change based adeptness or substrate choice

Attempt 2: Order chosen using randomization

- **Plan**: the chemist using a system to randomize the order of the 20 runs
- **Impact**: the chemist's change in adeptness will not benefit only one substrate
- **Result**: everything's ok!

Dealing With Extraneous Variables, cont.

Technique 3: Replication

- def: the process of repeating a run of the experiment more than once for each combination of experimental variables.
- results on the first run should be similar to the results on the second run if no experimental variables are changed - changes are the result of run-specific extraneous variables.
- after repeating multiple times, impact of run-specific effects average out
- neat: replication is strongly connected to the concept of reproducibility - the results I get should be similar to the results you get

In Summary

The main point is this: an effective experiment is **designed** to account for the environmental conditions that could influence our response. Doing this takes a lot of planning.

Section 2.4

Some Common Experimental Plans

Recap

Common Plans

Designing Experiments

Common experimental designs

Completely randomized experiment

There are many subtleties that enter into the planning of an effective experiment. There are some standard "skeletons" of plans that can help with planning an experiment.

Completely randomized experiment

A **Completely randomized experiment** is one in which all experimental variables are of primary interest (i.e. none are included only for purposes of blocking), and randomization is used at every possible point of choosing the experimental protocol.

Recap

Common Plans

Designing Experiments

Common experimental designs

Randomized complete block experiment

Randomized complete block experiment

A **randomized complete block experiment** is one in which at least one experimental variable is a blocking factor (not of primary interest to the investigator); and within each block, every setting of the primary experimental variables appears at least once; and randomization is employed at all possible points where the exact experimental protocol is determined.

Recap

Common Plans

Designing Experiments

Common experimental designs

Example:[Glass restrengthening]

Boyer, Millis, and Schiber studied the restrengthening of damaged glass through etching. They investigated the effects of the concentration of hydroflouric acid in etching bath and the time spent in the etching bath on the resulting strength of damaged glass rods.

Strengths were measured using a three-point bending method.

- Experimental variable(s): ① Concentration of hydroflouric acid ② time.
- Response variable(s): strength of damaged glass.

Recap

Common Plans

Designing Experiments

Common experimental designs

Example:[Glass restrengthening]

A $\boxed{3 \times 3}$ factorial experiment is run with the levels of concentration being 50%, 75%, and 100% HF and the levels of time being 30 sec, 60 sec, 120 sec. 18 damaged rods were allocated - two apiece - to each of the nine treatment combinations for testing.

9 treatments (all possible combinations of levels of experimental variable)

This was done at random by labeling the rods 01-18, placing numbered slips of paper in a hat, mixing, drawing two out for 30 sec. and 50%, then drawing two out for 30 sec. and 75%, etc. The slips of paper were also used to determine the order of testing and the order of damaging the rods.

Completely randomized design or Randomized complete block design?

Concentration of Hydroflouric acid

Time Spent

	50%	75%	100%
30 sec	J_{11} J_{12}		
60 sec			
120 sec			

Recap

Common Plans

Designing Experiments

Common experimental designs

Example:[Chemical purity, cont'd] (slide 26)

Assume time of day is an extraneous variable.

Completely randomized design: ?

Randomized complete block design: ?

Ractant A has two levels (a_1, a_2)

Reactant B has two levels (b_1, b_2)

Assume we collect 4 data points for AM shift & 4 data points for PM shift.

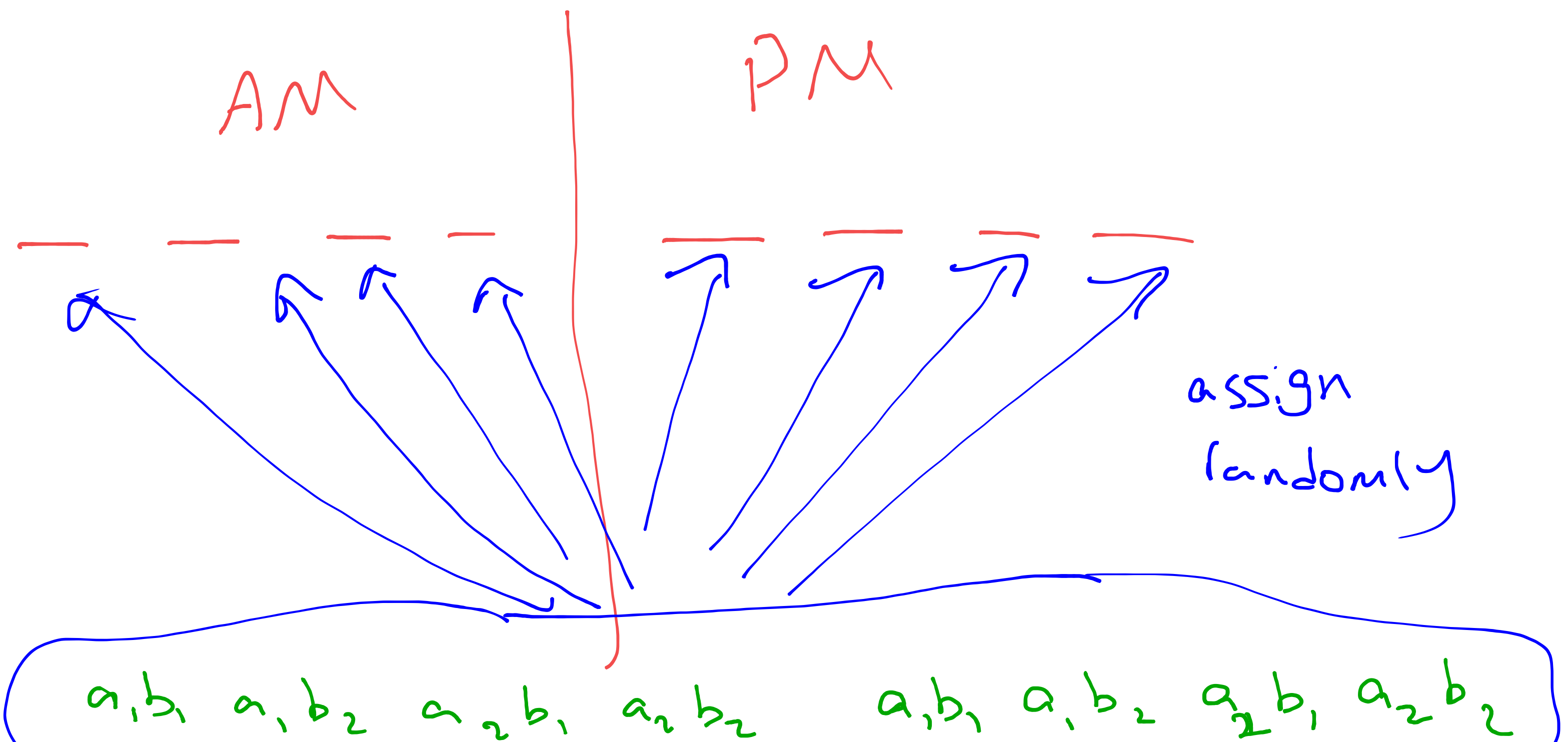
① what are all possible combinations of the two experimental variables?

* $a_1 b_1$

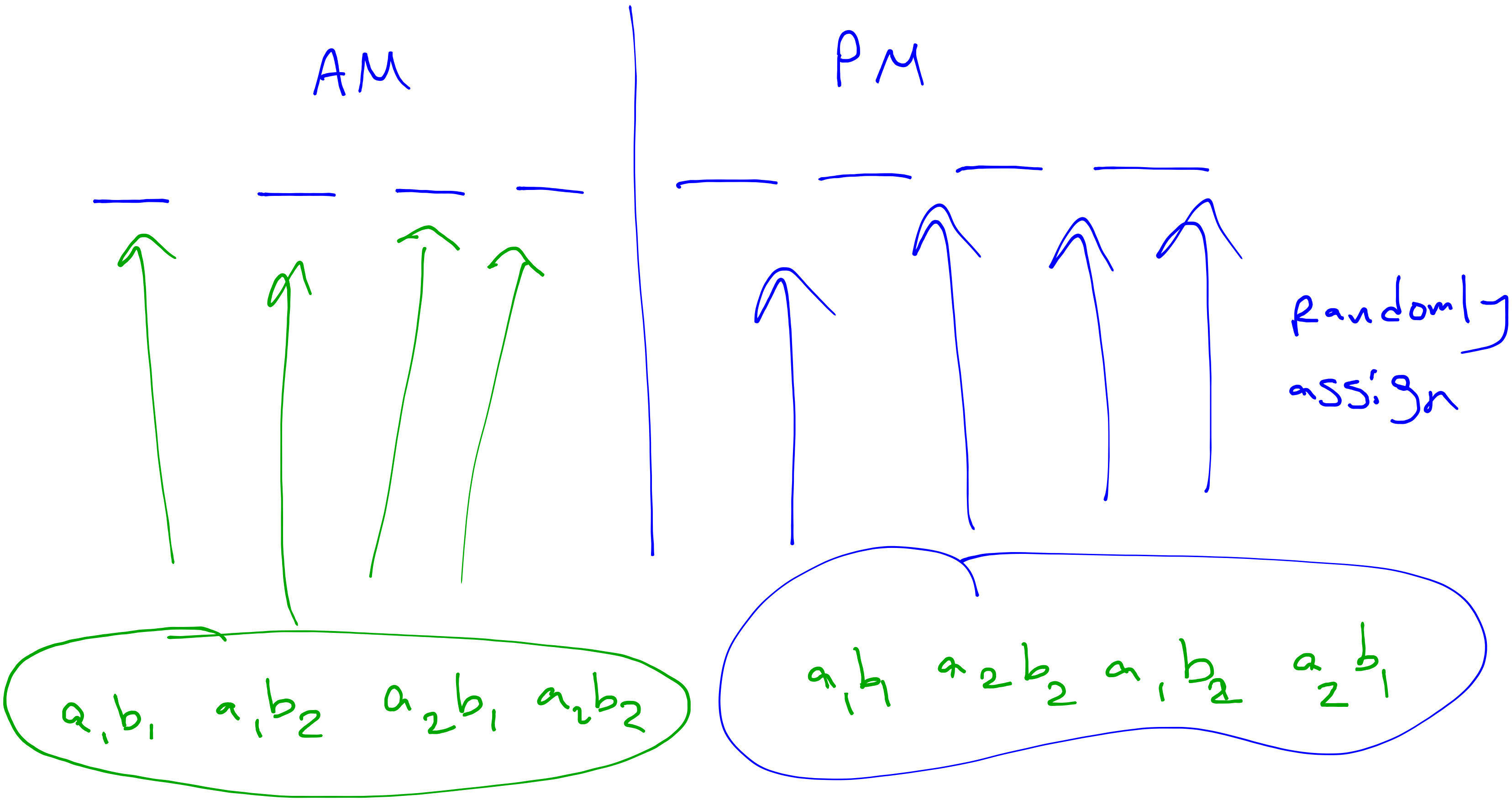
$a_1 b_2$

$a_2 b_1$ $\quad$ $a_2 b_2$

- Completely Randomized design:



- Randomized Completely Blocked designs



Section 2.5

Preparing to Collect Engineering Data

Read Independently