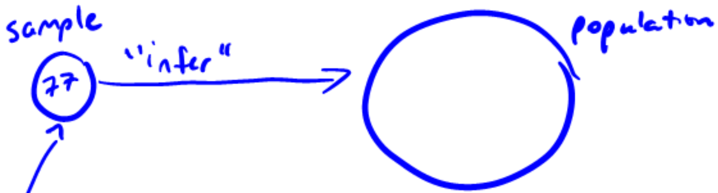
2 Data collection



Data collection is one of the most important parts of engineering statistics. If collected properly, data can make formal inferences easy to complete and easy to understand. On the other hand, if data is collected poorly, it can become nearly impossible to salvage a badly designed study and gain insights.

This chapter covers the general principles of data collection, ideas for effective experimentation, and examples of common experimental setups.

2.1 Sampling

Q: The most common question engineers ask about data collection is

How many observations do I need?

A: The answer depends on the variation in response that one expects.

Often we want to answer a question (conduct a study) about an identifiable, concrete population of items, but we want to use a **sample** to represent this (typically) much larger population.

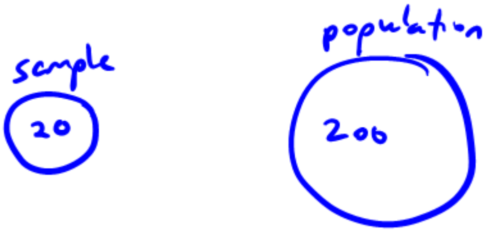
Why? save time, save money, maybe measurements destroy the sample, impossible (population is too large / unattainable).

Example 2.1. Measuring some characteristics of a sample of 20 electrical components (note: this is one sample with 20 units; the sample size is 20) from an incoming lot of 200.

If a sample is to be used to stand for a population, how that sample is chosen becomes very important.

A sample should

be representative of the population



2.1.1 Systematic and judgement based methods

Definition 2.1. In *systematic sampling*, create a list of every member of the population. From the list, randomly select the first sample element from the first k elements on the population list. Thereafter, we select every k^{th} element on the list.

Disadvantage: It can fail when cyclical patterns are present.

Definition 2.2. In *judgement-based sampling*, select based on the opinion of an expert.

Disadvantage: subject to unconscious/conscious bias and preconceptions.

2.1.2 Simple random sampling

(SRS)

Definition 2.3. A *simple random sample of size* n from a population is a sample selected in such a manner that every collection of n items in the population is a priori equally likely to compose the sample. equal chance of being selected.

Example 2.2. A statistics instructor wanted to know how many hours per week her students spend watching cat videos on YouTube. Rather than asking each one of them, she puts all of their names in a hat and draws out 10. This is a simple random sample of size 10.

Steps to randomly sample mechanically:

1. Let M be the number of digits in the number N, where N is the population size.
2. Give each member of the population an M-digit label.
3. Move through the table of random digits (Table B.1) from left to right, top to bottom, selecting population members for the sample when you encounter their indices (ignoring indices that have already been chosen) until you have selected n units for the sample.

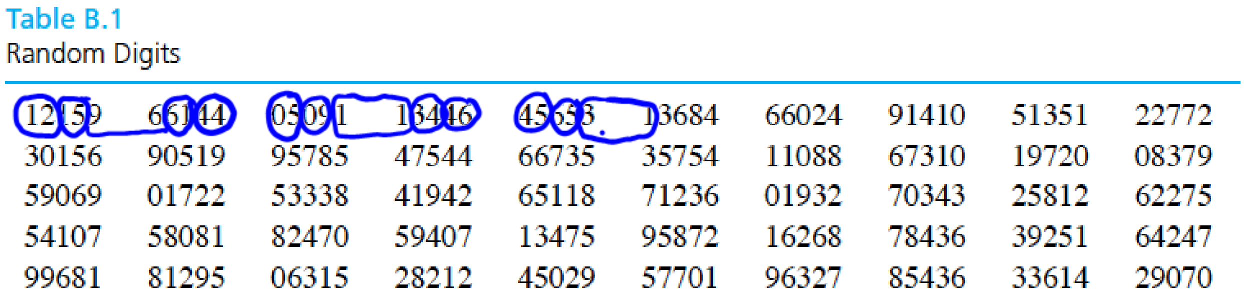


Table B.1
Random Digits

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

Example 2.3. Take a simple random sample of 12 units of pig iron out of a shipment of 90 units.

$N = 90$ $n=12$

$M = 2$

12, 15, 61, 44, 05, 09, 11, 34, 46, 65, 31

Alternatively: Use a computer.

JMP → free for this class

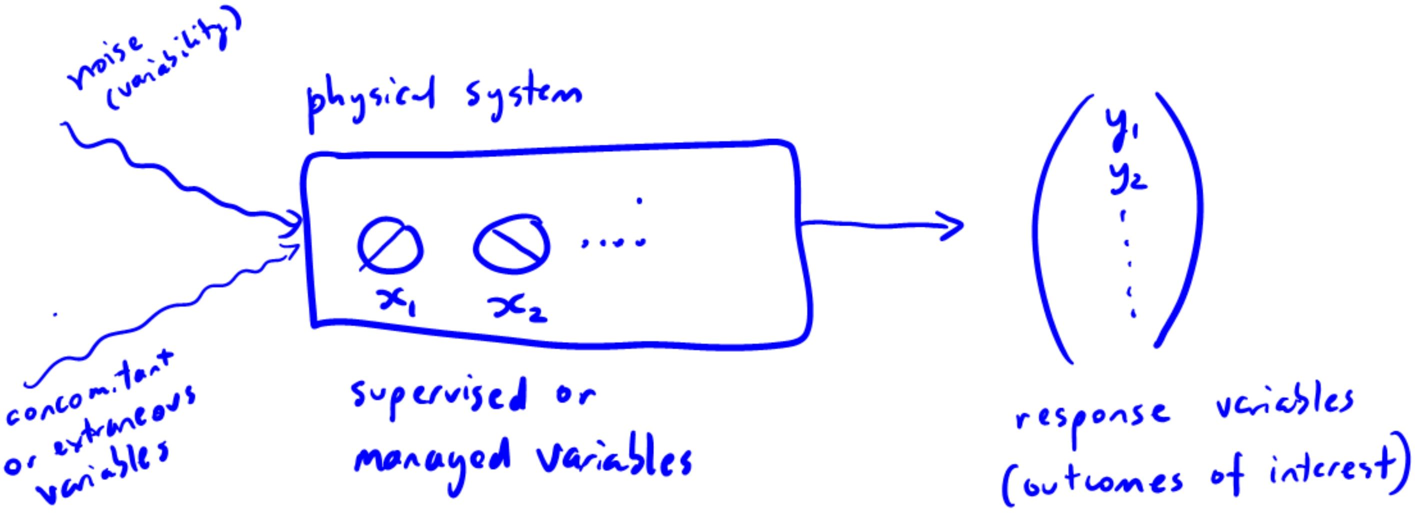
R → free software, open source

Excel ... etc.

2.2 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:


noise (variability)
physical system
x_1
x_2
y_1
y_2
concomitant or extraneous variables
supervised or managed variables
response variables (outcomes of interest)

Example 2.4 (Chemical purity)**.** Suppose you want to know about the effect of two different reactants (A and B) on the purity response 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 $b2$).

factors w/ 2 levels

2.2.1 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.

Definition 2.4. A *response variable* in an experiment is one that is monitored as characterizing system performance/behavior. outcome of interest

Definition 2.5. 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*.

Definition 2.6. 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. observed, but not changed and it isn't the quantity of interest (response)

Example 2.5 (Chemical purity, cont'd)**.** What are the response variables, controlled variables, experimental variables, and concomitant variables?

response: purity of chemical

controlled variables: mixing speed, batch size

experimental variables: reactant A
reactant B

concomitant variables: air temperature
humidity
time of day

Example 2.6 (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 ① full/complete factorial study for this experiment? ② What are the response variables, ③ controlled variables, ④ experimental variables, ⑤ and concomitant variables?

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

controlled variables: drying time
pressure

experimatal variables: wood type (1, 2, 3)
glue type (A, B, C)

concomitant variables: moisture content

Factorial study (3×3)

wood type	glue type
1	A
1	B
1	C
2	A
2	B
2	C
3	A
3	B
3	C

2.2.2 Extraneous variables

Definition 2.7. *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:

1. treat them as controlled variables (hard to extend results)
2. handle them as experimental variables and create several homogeneous environments to compare levels of primary experimental variables (blocking)
3. randomization (sometimes extraneous variables are overlooked)

Definition 2.8. 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.

Definition 2.9. *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.

e.g. assigning students to two groups by drawing names out of a hat.

Example 2.7 (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?" 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. What are the response variables, controlled variables, experimental variables, and extraneous variables? How would you use handle the extraneous variable (three ways)?

response: thrust face runout

controlled variables: ——

experimental variables: loading method (hang vs. lay)

extraneous variable: gear type

① controlled variable

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

② Blocking

If there are two types of gears, apply both loading methods (hang/lay) to both types of gears.

③ Randomization

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

Advice: Control and block for what you can and randomize for the rest.

2.2.3 Some more key issues of data collection

1. Comparative study Need a point of reference

   e.g. studying the strength of a new alloy, may need the strength of the existing alloy.

★ 2. Replication Need evidence that a study is repeatable or reproducible (i.e. results are not from chance or a mistake)

   ⇒ under the same settings, need to collect data more than once.

3. Allocation of resources Need to plan ahead

   Spend your resources (time/money/lab space) sequentially (i.e. expand your experiment gradually if possible) to get preliminary results.

   If data variability is high ⇒ need more data

2.3 Common experimental designs

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.

Definition 2.10. 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.

Definition 2.11. A *randomized complete block experiment* is one in which at least one experimental variable is a block 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.

Example 2.8 (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.

A 3 × 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. 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.

replication

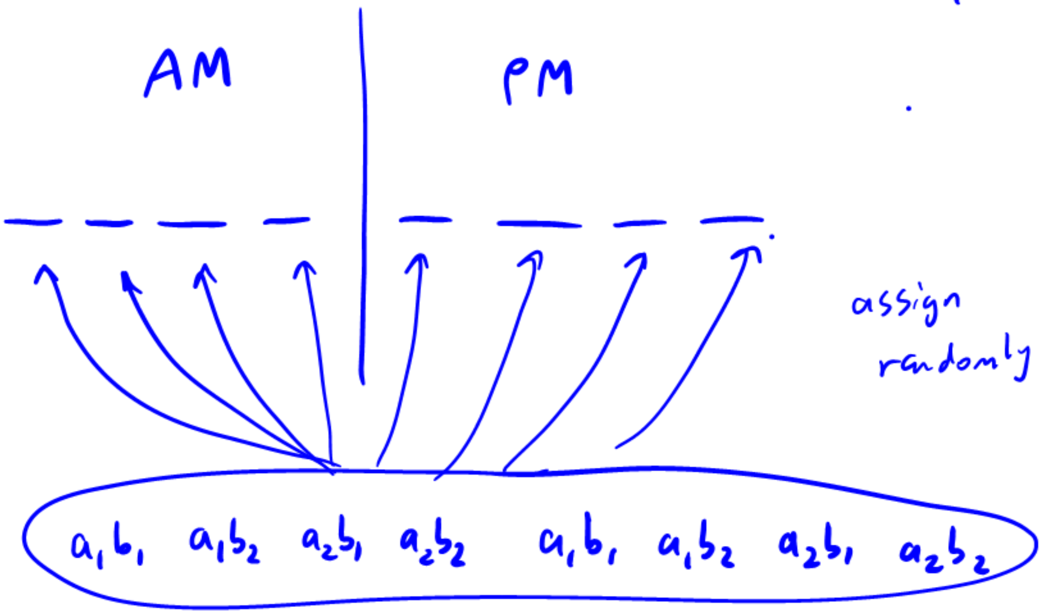
[✓ **Completely randomized design**] or **Randomized complete block design?**

Reactant A has 2 levels (a_1 and a_2)
Reactant B has 2 levels (b_1 and b_2] 2×2 = 4 combinations

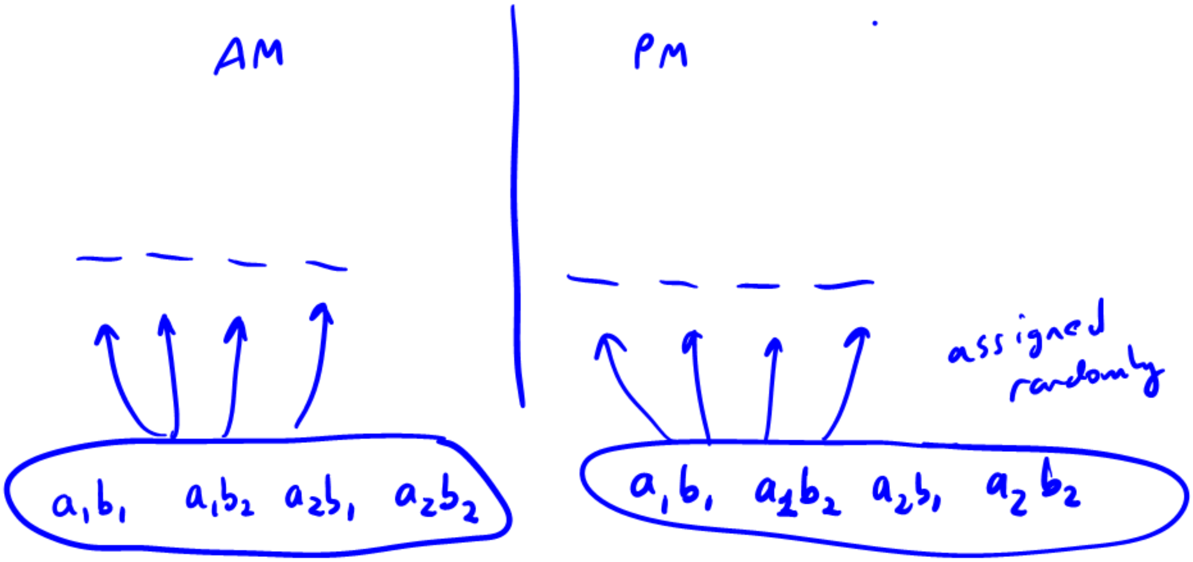
Suppose we need 2 batches of all 4 combinations (replication).

Example 2.9 (Chemical purity, cont'd). Assume time of day is an extraneous variable.

Completely randomized design: collect 4 data points for AM shift and 4 data points for PM shift


AM
PM
assign randomly
a_1b_1 a_1b_2 a_2b_1 a_2b_2 a_1b_1 a_1b_2 a_2b_1 a_2b_2

Randomized complete block design:


AM
PM
assigned randomly
a_1b_1 a_1b_2 a_2b_1 a_2b_2
a_1b_1 a_1b_2 a_2b_1 a_2b_2