STAT 305: Chapter 2

Data Collection

Amin Shirazi

Course page:
ashirazist.github.io/stat305.github.io

Quick Recap: Populations and Samples

Recap

Making Generalizations

When performing an experiment or gathering data in an observational study, the (main?) goal is to take the information you learn and apply it *outside* of your experiment - i.e., to make *generalizations*. For instance, we may wish to

- describe a relationship between two groups when we do not have the time or ability to gather information from from each member of the two groups,
- use the results of our experiment to predict the outcome of a scenario that has not yet occured,
- explain what part of a process are making the largest contribution to inconsistent results, and so on.

Our ability to make *valid* generalizations heavily depends on the validity of two parts of the study's setup: our **population** and our **sample**.

Recap

Populations

def: A **population** is the entire group of objects about which one wishes to gather information in a statistical study.

Important: A study's population should be *clearly described* - there should be no question about which objects are in the population and which are not. If a study's population is *not clearly described*, then regardless of how well you execute the mechanics of the study, you will be left with the following conclusion:

> In conclusion, after performing this study we can safely say that our results can be applied to ???

Quick question:

If our goal is to make statements about a population, why don't we just study the population?

Recap

Populations

- Logistical issues
- Timeless
- Expensive
- Destructive to the objects under study
- ...

Recap

Samples

def: A **sample** is the group of objects on which one actually gathers data.

These should be members of the population about which one wishes to gather information in a statistical study.

Recap

Getting Samples

The purpose of the sample is to be a representation of the population that can actually be studied in depth. Thus, the goal when gathering the sample is to make sure that there is no question that the sample actually does represent the population. A good sampling tecnique gives your study a indisputable connection between the sample and the population.

The gold standard of sampling methods is **Simple Random Sampling**. Using SRS, every possible sample of the same size has the same likelihood of being the sample used in the study.

Recap

Getting Samples

However, real-world physical constraints may make simple random essentially impossible. In other words, there are "possible samples" from our population that are more likely to used in our study than others. The degree to which our study makes using some samples more likely than others is called **bias**.

In this case, we may have to make (or ask others to make) additional assumptions in order to minimize *the impact of the biased sampling* and still connect the sample we have with the population we are interested in.

Recap

Getting Samples

Example: In a study of lifetime of lightbulbs, we took 100 consecutive lightbulbs off the factory line and measured their effective lifetime. We found that approximately 95% of lightbulbs survived 2,000 hours of use. We determine that 95% of the lightbulbs produced by our plant will survive 2,000 of use.

- population:
- sample:
- hidden assumption connecting the sample and the population:
- highly biased?:

Recap

Getting Samples

Example: In a study of video games effects on emotions, 200 college students were asked how often they played video games and how often they felt angry. The researches found a strong positive correlation between the number of hours spent playing video games and the number of times the student felt anger. They concluded that video games led to increased anger.

- population:
- sample:
- hidden assumption connecting the sample and the population:
- highly biased?:

Recap

Getting Samples

Example: As part of a study of the health of animals on campus, a field worker set traps and captured 200 squirrels. Once captured, a squirrel was measured and weighed, had it's age estimated, and blood was drawn to test for disease. After being held for a day, the squirrel was chipped and returned to the wild. The researchers reported that squirrels on campus were underweight.

- population:
- sample:
- hidden assumption connecting the sample and the population:
- highly biased?:

Quick Overview of Chapter 2

What We Need To Know

Section 2.1:

General Principles in the Collection of Engineering Data

(Read on Your Own)

Section 2.2:

Sampling in Enumerative Studies

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

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

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

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"

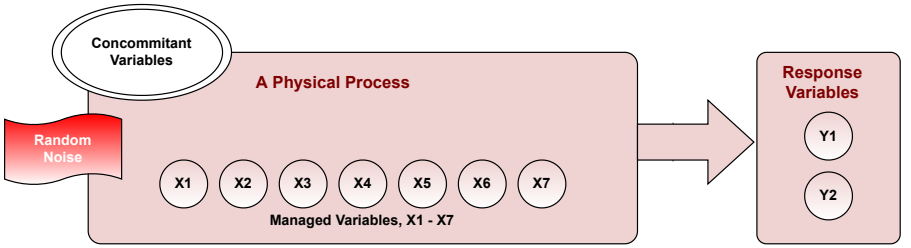
sample(1:25, 3)

[1] 20 19 24

Section 2.3

Principles for Effective Experimentation

The Ideal Experiment Structure



A few terms to help us make sense of the natural complexity of characteristics influencing system performance:

- **Response variable**: the characteristic indicating system performance which is being monitored.
- **Supervised** or **managed variable**: the characteristics of the system that the investigator can control.
  - **Controlled variable**: a supervised variable that is held constant throughout the experiment.
  - **Experimental variable**: a supervised variable that is given several different settings during the experiment.

The Ideal Experiment Structure, cont.

- **Blocking variables**: characteristics of the system that can be manipulated to create homogeneous environments within which to compare the effects of the primary experimental variables.
    - This is essentially extending the idea of control variables - we just create several environments with different controls.
    - How to recognize - the comparisons are not made comparing results from one block to results from another but instead comparing results inside a block.
- A **block** of objects under study is a homogeneous group where different levels of experimental variables can be applied and compared in a relatively uniform environment.

The Ideal Experiment Structure, cont.

Example on blocking

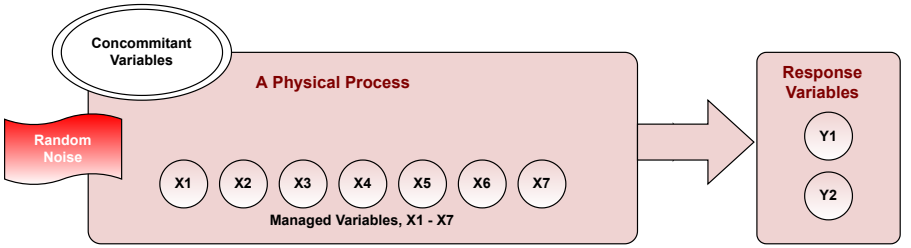
Back to machine parts testing example in lecture 2, in a shipment of 5000 parts. The goal is to verify that the parts are as strong as we anticipated. Assume the hardness of the parts are pre-specified by the seller and come in separate batches. Assume there are four levels of hardness (3.0, 3.2, 3.4, 3.6)

- We might want to test the Rockwell Hardness test across specimens of various hardness levels.
  - We then take samples from each batch
  - Hardness level is the blocking variable

The Ideal Experiment Structure, cont.

- **Concomitant variable**: characteristics that are observed but are not managed or responses. Could be influenced by either experimental variables or unobserved causes. May or may not have an influence on the response.

The Ideal Experiment Structure, cont.



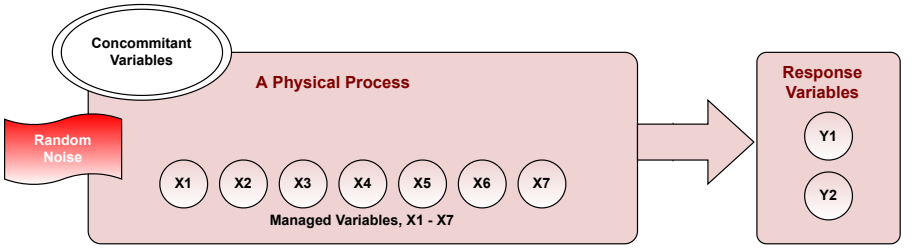
Extraneous Variables

There are lots characteristics that could influence the response but are not of primary interest to the experimenter. For instance,

- The experimenter could be unaware of their importance,
- There may be no way to control them in the experimental setting,
- There may be no way to control them outside of the experimental setting,

However, if we ignore them completely, their effect won't just disappear - it could ruin our experiment.

The Ideal Experiment Structure, cont.

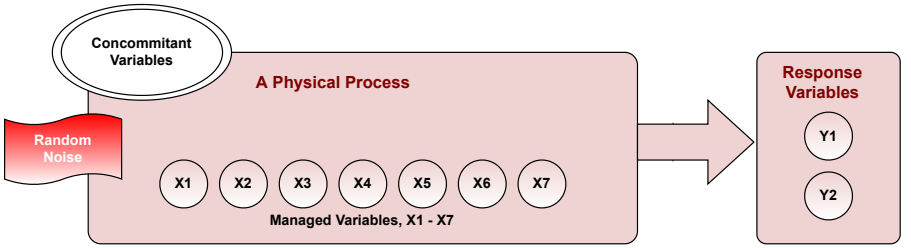


There are two common ways to attempt to account for these effects:

1. **Blocking**: Treat the extraneous variables as blocking variables

2. **Randomization**: assign runs of the experiment to the different levels of the extraneous variables randomly, with the hope that it balances out in the end.

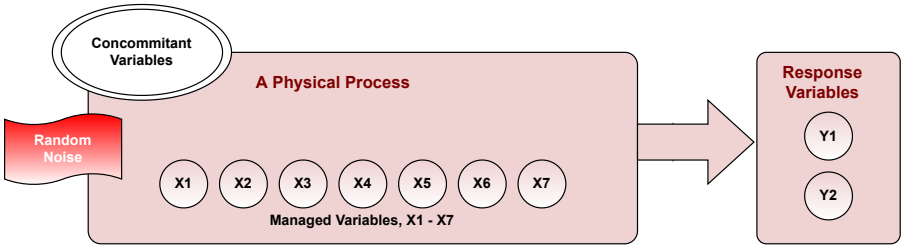
The Ideal Experiment Structure, cont.



Example: In the heat treating example (example 1in Chapter 1), Loading method was the primary experimental variable. Extraneous variables **Steel Heat** and **Machining History** were controlled by experimenting on 78 gears from the same heat code, machined as a lot. Then 78 gears were randomly assigned to two groups (one group were laid and the other were hung from a bar). Without randomization, unintended changes in measurement techniqes would appear in the data as differences between the two loading methods.

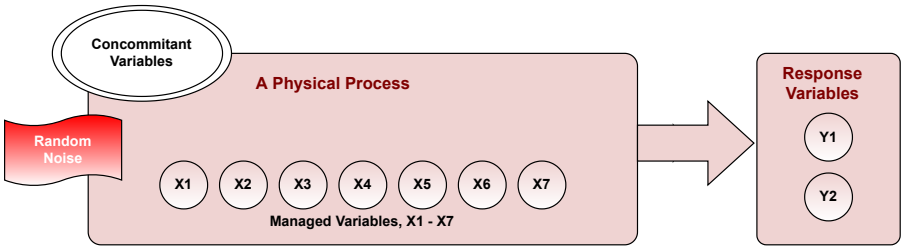
Practice with measurement equipment might increase the precision and make the later measurements to be more uniform than the early ones.

The Ideal Experiment Structure, cont.



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

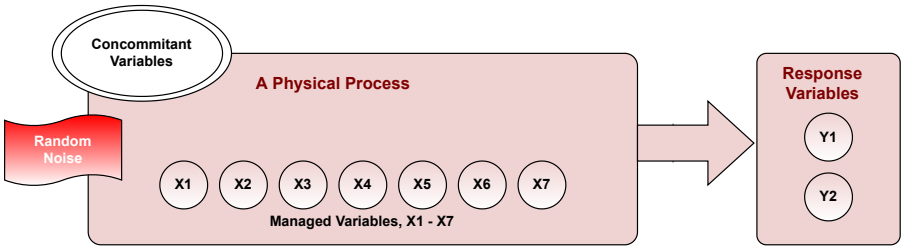
The Ideal Experiment Structure, cont.



Comparative study: Need a valid point of reference - so if we want to know if the new is better than the old, you better try to get some comparable data on the old as well. Without comparison, there is no firm basis on which to say that any effects observed come from the new conditions under study rather than from unexpected extraneous sources.

Repitition: Multiple responses measured from the same conditions.

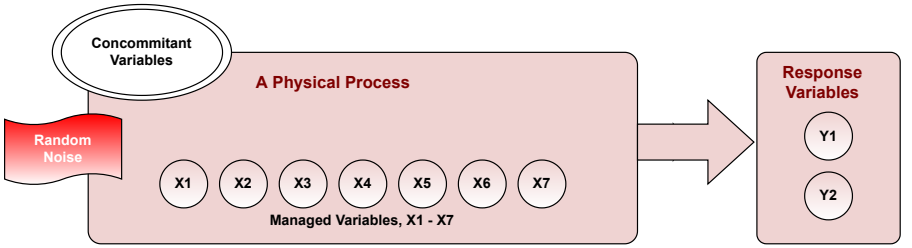
The Ideal Experiment Structure, cont.



Discussion: Example 7, pg. 39

- Three types of wood and three types of glue, two students sought to investigate joint strength.
- Issues: *drying time* and *pressure applied* during drying also important; smooth vs. rough wood; wood species have different moisture contents; the experiment is performed over two time periods.
- Approach: all wood/glue combinations dried 90 minutes with same pressure applied, moisture content of wood type measured before gluing.

The Ideal Experiment Structure, cont.



Discussion: Example 7, pg. 39

How do you measure moisture content? Cut the piece in half, record original weight and dried weight.

- Supervised(Managed) variables: wood, glue, time, pressure
  - controlled: time, pressure
  - experimental: wood, glue
- Concomitant: moisture

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.

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.

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

Common Experimental Plans

Designing Experiments

Experiments don't just happen, they have to be planned out ahead of time.

With **replication**, **randomization**, **blocking** we have introduced some common techniques used in planning (or **designing**) an experiment.

Common combinations of these techniques can be used as frameworks around which we can build our real-life experiments.

Some of these combinations have become so useful that we named them to make explaining the experimental design easier.

Design 1: Completely Randomized Experiments

Description

- All experimental variables are of primary interest (no controls, no blocking)
- Randomization used at every possible point where we choose how to treat the experimental units

Example

- Example 12 (page 51): Golf ball drive distances study by G.Gronberg
  - 30 balls in total. 10 balls each at 3 different compressions (80, 90, and 100)
  - How would you design an experiment to determine which type can be driven the furthest?

The levels of the experimental factor (compressions) are an intrinsic property of the balls. There is no way to randomly divide the 30 balls into three groups and "apply" the treatment levels 80, 90 and 100. Instead, the randomization could be the choice of an order for hitting the 30 test balls.

Design 2: Randomized Complete Block Experiments

Description

- At least on factor is a blocking variable
- "Blocks" are created by combinations of the levels of the blocking variables
- Within each block, every combination of the primary experimental variables are used at least once
- Randomization is used where possible

Examples

- Example 12 (page 51): Golf ball drive distances study by G.Gronberg
  - Gronberg didn't drive the golf balls all at the same time
  - instead, he personally drove 10 golf balls from each compression each night, 6 nights a week, for 3 weeks.
  - The block (days) account for possible changes over time in his physical condition and skill level and environmental conditions.

Design 3: Randomized Incomplete Block Experiments

Description

- At least one factor is a blocking variable
- "Blocks" are created by combinations of the levels of the blocking variables
- Within each block, some combinations (but not all) of the primary experimental variables are used at least once while others are not used at all
- Randomization is used where possible

How does this work??

- There are some clever ways to assign the combinations of the primary experimental variable levels to the blocks that allow us to get "more" out of the data we collect.
- How these block assignments come about gets explained in Chapter 8.

Section 2.5

Preparing to Collect Engineering Data

Read Independently