Appendix - Blocking and Stratification

Contents

This course covers two fundamental concepts in causal inference: **blocking** in **experimental studies**, and **stratification**. Now you might wonder:

> *What is the difference between **blocking** and **stratification**?*
>
> *Is it that **stratification** is a general term for splitting similar data points into groups, and **blocking** is what we do when we set up an experiment by stratifying?*
>
> *Because I thought blocking was something done in the experimental design stage, and stratification is used in an observational study to deal with confounding.*

Let us dig into some clarifying ideas.

Relationship between Blocking and Stratification

A few terms are related to one another: **stratification**, **blocking**, **randomization**, and **confounding**.

The first thing to remember is that this course (**DSCI 554**) is not about running a Machine Learning model on already collected data. Instead, it is about **planning**, **designing**, and **analyzing** an **experimental** or **observational study** to draw **causal** inferences using a certain factor-type variable X on a response of interest Y (either discrete or continuous).

In the case of experimental studies, the **design** aspect of **DSCI 554** is why so many new statistical terms are brought up. If we do not care about the experimental design, all we are left with is the end dataset. Furthermore, we will lose the vital leverage we have using treatment randomization before collecting our experimental data.

Why is the vital leverage from the experimental design so important? So our model could provide more precise and accurate estimates of causal interest, while dealing with non-observed variables in our experimental units (**the so-called confounders!**).

Now, let us move on to the following definitions:

- **Stratification.** It refers to specific attributes of the **population of interest**. This matter can be something that does not really change about the individuals in the population (e.g., *day* versus *night* visits to a website) or something we, as the researchers, can define (e.g., taking the age of individuals and putting them into age groups).
- **Blocking.** In an **experimental study**, if we know there is a variable related to the experimental variables we care about (but not a variable we want to study per se!), then we can account for this variable with blocking. It has two steps to it:
    1. **Stratify experimental units** by using the same strata from the population. Each stratum will be a **block**, and this variable is your **blocking variable**.
    2. **Randomize from the population into your individual blocks.** How would we do this in terms of **sampling from the population of interest**? We stratify our population and then **randomly sample from each stratum into its corresponding block**.

Once you have our data in separate blocks, we randomize the sampled experimental units into the **treatments** (**control**, **experimental treatment 1**, **experimental treatment 2**, etc.). Then, the experimental intervention will proceed based on these treatments. Finally, we collect the data and analyze it.

The diagram below shows this process as an example involving a clinical trial (i.e., an experiment).

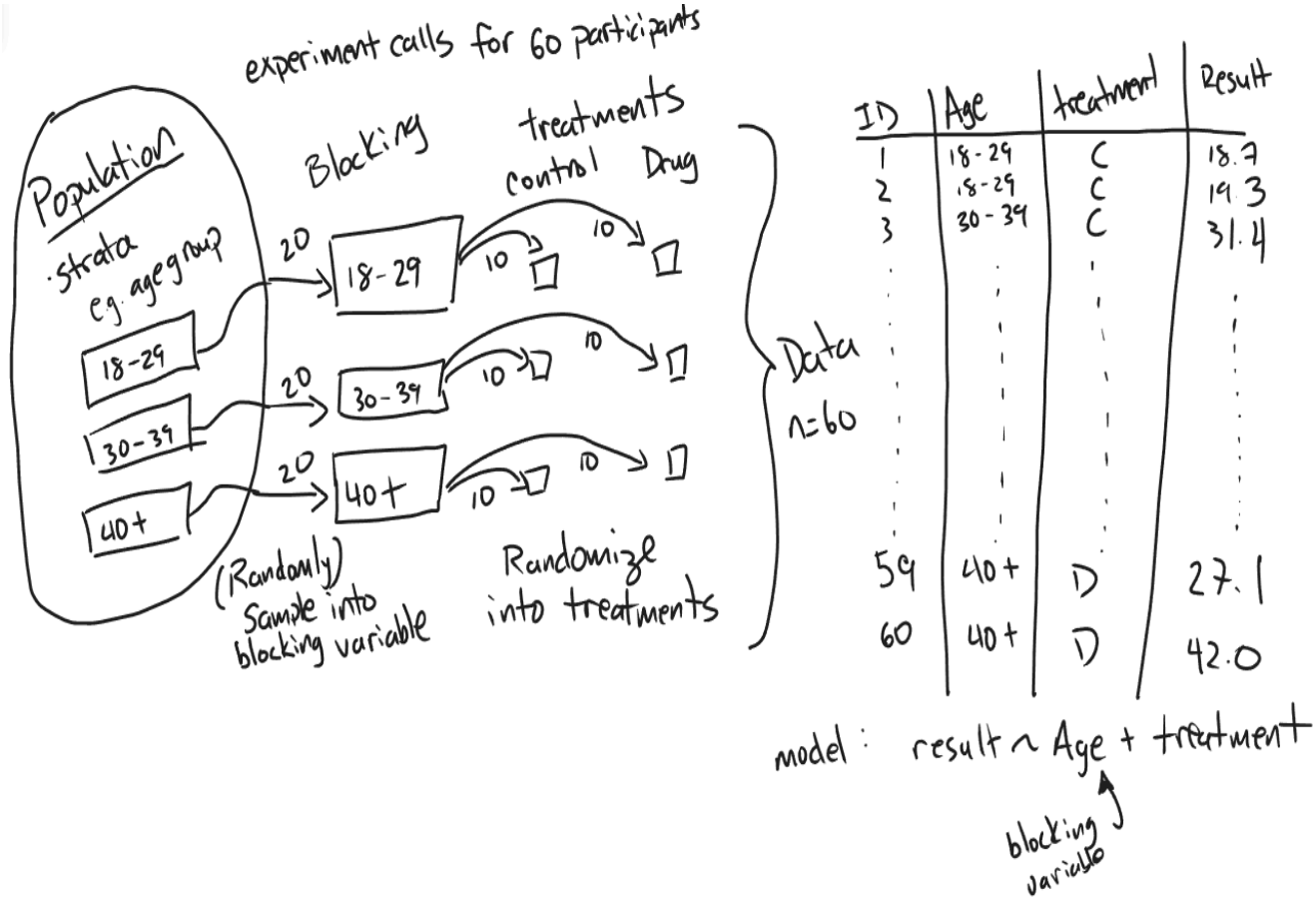

Fig. 5 Relationship between experimental blocking and population strata (credit: Daniel Chen).

Why Blocking might be confusing?

Because the process of **blocking** or having a **blocking variable** is a specific term used when **designing experiments**. So, for example, if we did not know about any population strata and were given just the collected experimental dataset, nothing about that dataset would tell you that blocking occurred (unless specified by the people who ran the experiment).

Assuming you did not know about any population strata and were given just the collected experimental dataset, you might imply that `age` (**for instance, in a given study**) may be a variable **you need to control** and fit a `result ~ age + treatment` model. **Under this assumption**, `age` will just be viewed as a **confounding variable**. Still, note that the `result ~ age + treatment` model will still have the **same systematic component** as if you were responsible for designing via `age` blocking with `treatment` randomization, running, and analyzing the whole experiment.

So, what is a confounder or counfounding variable?

It is a variable that your main study is not concerned with but is related to your experimental variables of interest.

Does not that sound like a blocking variable? Yes, but a blocking variable will be considered since the early stages of the experimental study (i.e., planning and design). Therefore, following up with the `result ~ age + treatment` model, **we would block by `age` and randomize `treatment` within the blocks in the design before running the experiment and collecting our data**. Then, finally, we would fit our `result ~ age + treatment` model.

Blocking Versus Stratification

Blocking is related to the **design of the experiment** that accounts for natural variability across subsets of experimental units that are homogeneous, while population **stratification** is more related to sampling methods (you can find more sampling information in these **DSCI 562 notes**).