Experimental Design

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

Statistics

Statistics is the science which deals with **the collection, the analysis, the visualization and the interpretation** of experimental data.

How data are collected . . .

- Random samplings
- Observational studies
- Experiments

Random Samplings (polls)

Definition

Random samplings allow to characterize the properties of a finite population without measuring all of its members.

Examples

- *Electoral polls*
- *Normal levels of cholesterol in the human population*
- *Characterization of a population of grapes, apples, wines, . . .*
- . . .

Observational Studies

Definition

Observational studies are designed with the objective of **identifying relationships** between the different properties of a conceptual population. The role of the experimenter is to perform the **selection of the sample**.

Examples

- *Is it true that people who eat more chocolate are more happy?*
- *The level of cholesterol of people eating more vegetables is lower ...*

Experiments

Definition

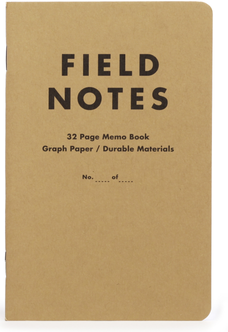
Experiments are designed with the objective of identifying **causal relations** between the properties of a conceptual population. The role of the experimenter is to **modify the conditions** to verify the presence of causal relationship between the observed properties.

Examples

- *If you eat more chocolate you will get happier*
- *If I drink more beer I'll get more sympathetic*
- ...

Important notes

- Causal relations can be assessed only in **experiments**
- This is really Galileian ;-)
- Experiments are impossible in many relevant fields like human health and ecology



Should we then give up on obtaining causal information there?

Mind the chocolate . . .

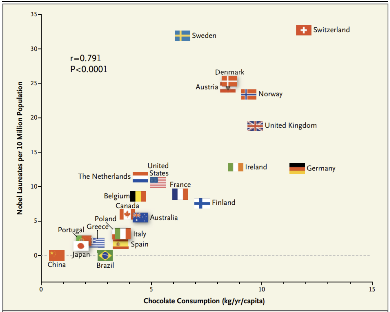


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

New England Journal of Medicine, 2012

" . . . Chocolate consumption enhances cognitive function, which is a sine qua non for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country . . . "

. . . and more!

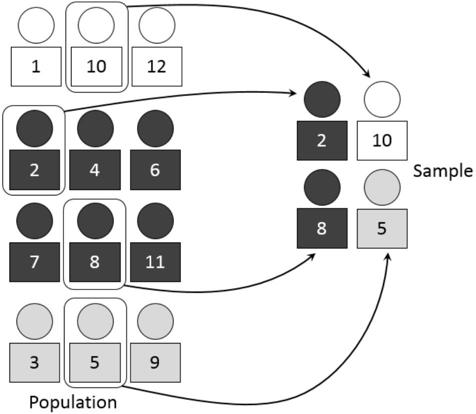
Planning a sampling

Key question: what is the best way to sample my population in a *representative* way?

- Do it *randomly* to avoid any intentional or unintentional bias (**Randomized Sampling**)
- Take into account known subpopulations and confounding factors (**Stratified Random Sampling**)
- The number of samples is determined by practical/economical considerations

In presence of known subpopulations stratified random sampling results in a more accurate characterization of the population

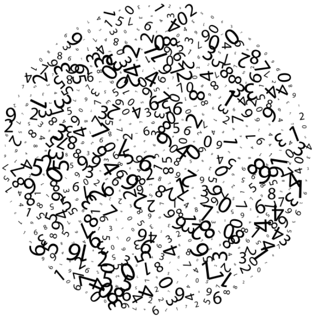
Stratified Random Sampling



Key idea: do it randomly

The most reasonable way to "smear out" the effects of unknown biases is to do everything **randomly**

Random is not a synonym of HAPHAZARD



Planning an observational study

Keep in mind . . .

Objective: get an useful and clear answer
Mean: start from a clear, useful and often simple question

- Identify the **sampling unit**
- Decide the number of samples (money, power, . . .)
- Define the conceptual population
- Sample it in a representative way
- Identify **confounding factors** and, if possible, **stratify** for them

Key idea: Sampling Unit

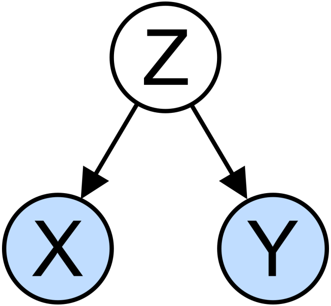
The smaller unit of a population which retains the properties we are interested into

- *Example: grapevine, leafs, infections ...*

Key idea: Confounding Factor

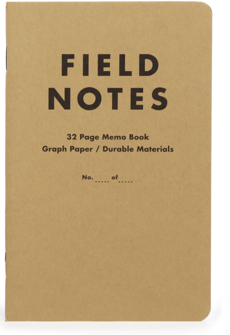
A variable that influences both the dependent variable and independent variable, causing a spurious association (*wikipedia*).

- *Smoke, cardiovascular disease, alcohol consumption*
- *Birth order (1st child, 2nd child, etc.), maternal age, Down Syndrome in the child*



Notes

- Some confounders can be controlled by careful sampling
    - Eg. *Age and Gender on the relation between happiness and chocolate*
- Some others are impossible to control
    - Eg. *Presence of chemical pollution in the water streams and altitude of sampling*



Planning one experiment

Keep in mind . . .

Objective: get an useful and clear answer
Mean: start from a clear, useful and often simple question

- What is my **experimental unit**?
- How many samples should I measure?
- What are the **potential sources of variability**?

Experimental Design

A strategy to assign the experimental units to the different *treatments* to optimize my capacity of inferring **causal relationships**

Control of unwanted sources of variability (technical/biological) to highlight the effects of the intervention

Key tool : Blocking

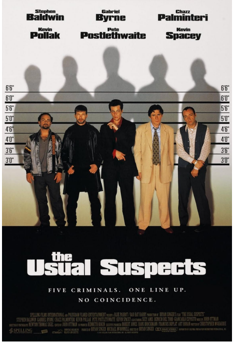
- Group experimental units in homogeneous groups (blocks)
- Study the variability inside the blocks
- Identify and subtract the variability across the blocks
- Blocking allows to *subtract* the difference between the blocks
- Blocks and study factors should be *orthogonal*

Block what you can; randomize what you cannot

Blocking is “better” than randomization because statistics allows to “subtract” the variability coming from the blocking factor

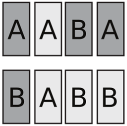
Examples of common blocking factors

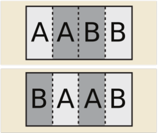
- Location
- Analytical batch
- Day
- Operator
- ...



Complete Randomized Design (CRD)

Randomized Complete Block Design (CRD)



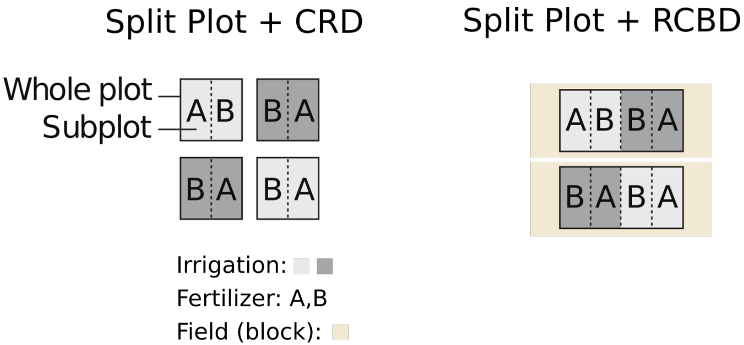


Irrigation:

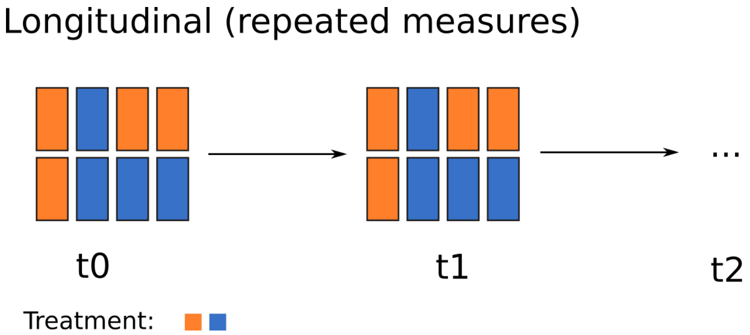
Fertilizer: A,B

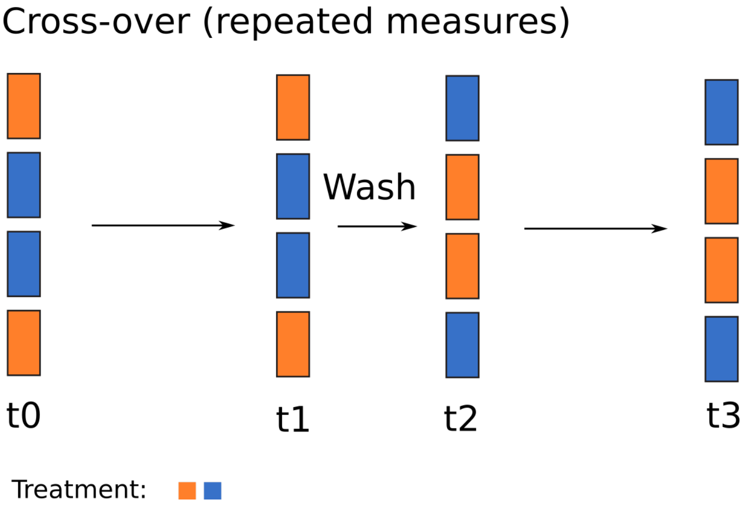
Field (block):

Split Plot Designs



Longitudinal Studies




Cross-over (repeated measures)
Wash
t0
t1
t2
t3
Treatment:

Notes

- Block as much as possible!
- Repeated measures are more "powerful" because each unit is the control of itself
- Crossovers can be tricky for the wash-out
- Repeated measures design are the key in presence of large variability in the population (e.g. plants in the field/greenhouse)