Esssentials of Applied Data Analysis
IPSA-USP Summer School 2018

Building a fake dataset

Leonardo Sangali Barone
leonardo.barone@usp.br

jan/18

Building a fake dataset of Fakeland citizens

Fakeland is a very stable democracy that helds presidential elections every 4 years. We are going to build the fake dataset of Fakeland individual citizens with information about their basic fake characteristics and fake political opinions/positions. We are going to do it in order to explore types of variables, to understand the link between probability rules and data analysis and to learn about random variables.

The variables that our fake dataset will contain are:

- *age*: age
- *sex*: sex
- *educ*: educational level
- *income*: montly income measured in fake money (FM$)
- *savings*: total fake money (FM$) in savings account
- *marrige*: marriage status (yes = married)
- *kids*: number of children
- *party*: party affiliation

- *turnout*: intention to vote in the next election
- *vote_history*: numbers of presidential elections that turned out since 2002 elections
- *economy*: opinion about the national economy performance
- *incumbent*: opinion about the incumbent president performance
- *candidate*: candidate of preference

Types of Variables

First, let's remember the types of variables:

A) Discrete - values are countable, even if infinite.

- Nominal - can be counted, but not ordered
- Ordinal - can be counted and ordered
- Integer - can be counted, ordered and are distance between categories are mathematically meaningful (yes, numbers with no decimals).

B) Continuous - Infinite set of values. Not countable (yes, whatever number there is, including $\sqrt{2}$ and π).

Types of Variables - Exercise

Look at the fake dataset you have in hands. Can you classify the variables by type?

Drawing data from fake distributions

To build the dataset one could simply type values in a spreadsheet. However, there is a more elegant strategy. Instead of filling a dataset randomly, I could determine, as if I were God, the **distributions of values in each variable** and sample from that distribution. In fact, the fake dataset you have in hands was built this way using *R Statistical Package*. You can find

the code that builds it in the course material.

By now you already know that in a sample space, each outcome is associated with a probability of occurence. Probability distributions – or probability models, as they are called sometimes – are simply the set of these probabilities. When I built Fakeland, I modeled the distribution of characteristics of it's citizens (Fakeland is my *model*).

In the next handout we will explore distributions of discrete and continuous random variables using our fake dataset. Keep in mind that the data in the sample comes from a "True" distribution for Fakeland citizens that I invented.