Representative Samples

José Bayoán Santiago Calderón

2018-11-20

1 Introduction

Our goal will be to learn how to generate representative samples for various covariates. This exercise will allow us to generate data that mirrors that of the US population for the civilian noninstitutional adult population (18 - 85 years old). Some covariates that might be of interest are `gender`, `age`, `race`, `weight`, and `height`.

1.1 Installation

Because PuMaS is still unregistered, you will need to give the Git repository in order to add the package. To do this, use the command `]add https://github.com/UMCTM/PuMaS.jl`. Doing it this way, PuMaS and its dependencies will install automatically. If one cannot authenticate for this command (since the repository is currently private!), then first clone the repository how you please, use `]add path/to/PuMaS.jl`, then then do `]build PuMaS`. Using the build command will download and install the dependencies.

1.2 Generating synthetic data

Data Sources:

```
Annual Estimates of the Resident Population by Sex, Single Year of Age, Race
Alone or in Combination, and Hispanic Origin for the United States: April 1,
2010 to July 1, 2017
Source: U.S. Census Bureau, Population Division
Release Date: June 2018
```

```
Lynn A. Blewett, Julia A. Rivera Drew, Risa Griffin, Miriam L. King, and
Kari C.W. Williams. IPUMS Health Surveys: National Health Interview Survey,
Version 6.3 [dataset]. Minneapolis, MN: IPUMS, 2018.
http://doi.org/10.18128/D070.V6.3
```

Gender and race are nominal variables which can be captured through a cross table with the join distributions. Age usually does not follow a typical distribution (i.e., demographic pyramids). We can use Census data to obtain the join distributions of gender and race for conditional on age which is a discrete variable.

Some housekeeping

```
using CSV, DataFrames, Distributions, StatsBase, Tables
```

We will read the data set and select the conditional probability array. The first dimension is the age in years (18 - 85), the second dimension is the race (White, Black, Asian, or Latino), and the third dimension is the gender (Male or Female).

```
joint_distribution = CSV.File("data/census.tsv",
                              allowmissing = :none,
                              delim = '\t') |>
                     DataFrame |>
                     (df -> cat(Matrix(df[df.Gender .== "Male", 3:end]),
                                Matrix(df[df.Gender .== "Female", 3:end]),
                                dims = 3))
size(joint_distribution)
```

```
(68, 4, 2)
```

The following functions apply the labels to each dimension and sample based on the joint distribution.

```
function gen_join_distribution(probabilities, output = Vector{Int}())
    d = ndims(probabilities)
    iszero(ndims(probabilities)) && return output
    o = sample(1:size(probabilities, d),
               sum.(selectdim.(Ref(probabilities),
                               d,
                               1:size(probabilities, d))) |>
               weights)
    push!(output, o)
    gen_join_distribution(selectdim(probabilities, d, o), output)
end
labs = [(:gender, ["Male", "Female"]),
        (:race, ["White", "Black", "Asian", "Latino"]),
        (:age, 18:85)]
function gen_gender_race_age(joint_distribution, labs)
    NamedTuple{Tuple(first.(labs))}(getindex.(last.(labs),
                                    gen_join_distribution(joint_distribution)))
end
```

```
gen_gender_race_age (generic function with 1 method)
```

We can use the various components to generate a sample of 1,000 observations that is representative of the US non-noninstitutional adult population (18 - 85 years old).

```
data = map(x -> gen_gender_race_age(joint_distribution, labs), 1:1e3) |>
       DataFrame |>
       categorical!
```

Using the National Health Interview Survey, we can estimate the conditional distribution of height and weight per gender and race. The normal distribution is a good fit for human heights for values around three standard deviations from the mean. Weight is log-normal distributed, but for the conditional distribution is easier to rely on a Weibull body mass index (BMI) distribution. The first step is to eliminate problematic outliers. For these purposes, observations were kept if BMI was between 15 (Very severely underweight) and 60 (Obese Class VI - Hyper Obese) and weight was between 30 kg and 140 kg. The BMI and weight implicitly apply a height outlier exclusion.