Surveys with Julia

Introduction to Survey.jl

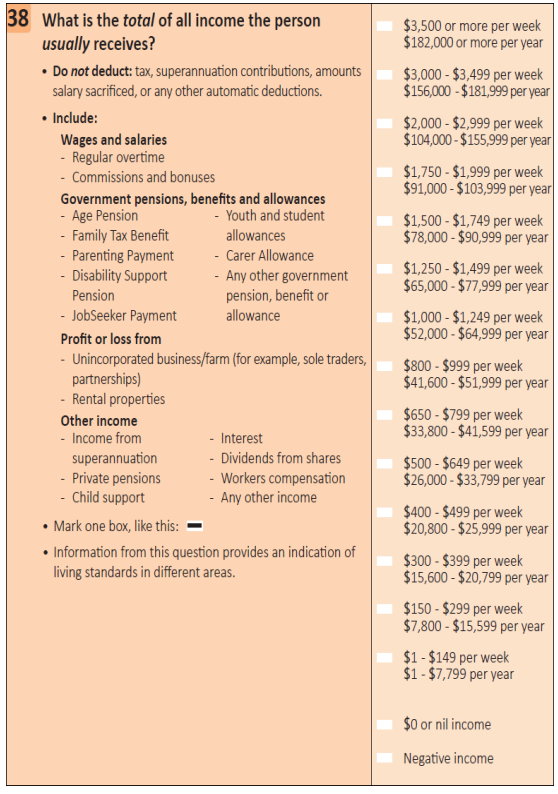
Shikhar Mishra

March 22, 2023

{X}KDR

What is complex survey analysis?

- Surveys are an empirical tool for social and behavioural analysis
- **Goal:** obtaining estimates for a large population by surveying a well selected subset
- In contrast to a **census**
- Techniques available for increasing precision and representation of the survey
  - Several types of survey "designs" and sampling methods



38 What is the *total* of all income the person *usually* receives?

- Do *not* deduct: tax, superannuation contributions, amounts salary sacrificed, or any other automatic deductions.
- Include:

Wages and salaries
- Regular overtime
- Commissions and bonuses

Government pensions, benefits and allowances
- Age Pension
- Family Tax Benefit
- Parenting Payment
- Disability Support Pension
- JobSeeker Payment
- Youth and student allowances
- Carer Allowance
- Any other government pension, benefit or allowance

Profit or loss from
- Unincorporated business/farm (for example, sole traders, partnerships)
- Rental properties

Other income
- Income from superannuation
- Private pensions
- Child support
- Interest
- Dividends from shares
- Workers compensation
- Any other income

- Mark one box, like this: ■
- Information from this question provides an indication of living standards in different areas.

- $3,500 or more per week / $182,000 or more per year
- $3,000 - $3,499 per week / $156,000 - $181,999 per year
- $2,000 - $2,999 per week / $104,000 - $155,999 per year
- $1,750 - $1,999 per week / $91,000 - $103,999 per year
- $1,500 - $1,749 per week / $78,000 - $90,999 per year
- $1,250 - $1,499 per week / $65,000 - $77,999 per year
- $1,000 - $1,249 per week / $52,000 - $64,999 per year
- $800 - $999 per week / $41,600 - $51,999 per year
- $650 - $799 per week / $33,800 - $41,599 per year
- $500 - $649 per week / $26,000 - $33,799 per year
- $400 - $499 per week / $20,800 - $25,999 per year
- $300 - $399 per week / $15,600 - $20,799 per year
- $150 - $299 per week / $7,800 - $15,599 per year
- $1 - $149 per week / $1 - $7,799 per year
- $0 or nil income
- Negative income

Some survey terminology

Weighting How many people does each respondent represent?

Strata Subgroups of the population known a priori eg. states, districts, gender. Strata info used to improve representation

Clusters Logistical constraints on survey sampling, can only visit n states, districts and suburbs

Why does a "survey analysis" package do?

- Computing summary statistics from a survey requires applying mathematical corrections and adjustments
  - eg. population mean is not as simple as arithmetic mean of a numeric vector
- Point estimation (relatively) easy
- Variance estimation is complex
- A "survey" package:
  - exposes an intuitive API to user
  - automatically applies formulae and corrections in background
- Eg. in Survey.jl for population mean (with SE) of a variable you can do `mean(:variable,data)`

Our engineering journey

- Users of R `survey` package
  - Benchmark for open-source complex survey analysis
  - CMIE CPHS, Prowess, NFHS etc
- R ‘survey‘ designed in early 2000’s for MB’s of data
  - slow for ”large” modern datasets and many class of simulation problems
  - eg. variance estimation using bootstrapping
  - Computation times upto few hours for simple summary statistics
- Real-world performance a key factor in development of `Survey.jl`

Why Julia for complex survey analysis

Performance **Expressivity** of R/Python meets **speed** of a systems language

Development Avoid "two-language problem". Survey researchers just want something that works great out of the box. Easy maintenance.

Community Several unmaterialised attempts to create survey analysis package. We received feedback and even contributing PRs. ▸ 1

Ecosystem Julia has substantial statistical computing abilities, with state of the art `DataFrames`, `Makie`, `Optim`, `Turing`, `Flux`, `LinearAlgebra` packages. `Survey` is complement to and complemented by the entire data ecosystem.

Survey.jl

An efficient computing framework for survey analysis

- Summary statistics - `mean`, `total`, `ratio`, and `quantile`
- Subpopulations / domain estimation for subsets of sample
- Variance estimation using (Rao-Wu) bootstrap
  - 1000 MC simulations for variance in Julia takes similar computation time as 50 simulations in R.
- Modularity - Can write custom replicate weighting algorithm
- Visualisations support for weighted scatter plots, histograms
- Tested and compared against R `survey`

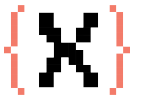
Getting started with Survey.jl GitHub and Documentation

Demo workflow

Import and load data

Survey: CMIE Consumer Pyramids Household Survey - Multistage stratified high frequency survey of Indian households

```
# Imports and housekeeping
...
# Connect to SQL server
conn = DBInterface.connect(MySQL.Connection, host, user,
  ↪  password; db = "hhd")
query = "SELECT RESPONSE_STATUS, STATE, HR, DISTRICT,
  ↪  STRATUM, PSU_ID, REGION_TYPE, FAMILY_SHIFTED, HH_ID,
  ↪  MONTH_SLOT, MONTH, TOTAL_INCOME , HH_WEIGHT_MS,
  ↪  HH_NON_RESPONSE_MS  FROM hh_income_monthly WHERE MONTH
  ↪  = 'Apr 2022' AND RESPONSE_STATUS = 'Accepted'"
# Pipe query output into DataFrame
df = DBInterface.execute(conn, query) |> DataFrame
```



Demo workflow

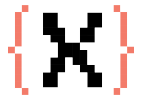
Create SurveyDesign

```
# Load df into survey design object
julia> CPHS_income = SurveyDesign(df, clusters = :HH_ID,
 ↪  strata = :STRATUM, weights = :HH_WEIGHT_MS)

SurveyDesign:
data: 123816×17 DataFrame
strata: STRATUM
    [HR 1_URBAN_S, HR 1_URBAN_S,  …  HR 110_RURAL_R]
cluster: HH_ID
    [5.3877505e7, 4.3406519e7  …  6.742216e7]
popsize: [8.81791619977e7  …  1.034108342135e8]
sampsize: [123816, 123816, 123816  …  123816]
weights: [712.1791, 712.1791, 712.1791  …  835.1977]
allprobs: [0.0014, 0.0014, 0.0014  …  0.0012]
```

Demo workflow

Create ReplicateDesign

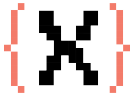
```
# Create replicate design using Rao-Wu bootstrap weights
julia> CPHS_income_bootstrap = bootweights(CPHS_income,
 ↪  replicates = 500)

ReplicateDesign:
data: 123816×517 DataFrame
strata: STRATUM
    [HR 1_URBAN_S, HR 1_URBAN_S  …  HR 102_URBAN_M]
cluster: HH_ID
    [1.0034716e7, 1.0190136e7  …  9.9842237e7]
popsize: [8.81791619977e7  …  2.92096840303e7]
sampsize: [123816, 123816, 123816  …  123816]
weights: [712.1791, 712.1791, 712.1791  …  235.912]
allprobs: [0.0014, 0.0014, 0.0014  …  0.0042]
replicates: 500
```



Demo workflow with CPHS

Calculate summary statistics

```
# Mean income (overall India)
julia> mean(:TOTAL_INCOME, CPHS_income_bootstrap)
1×2 DataFrame
 Row   mean      SE
       Float64   Float64

   1   23870.2   81.8377

# Total income by homogenous regions (Subpopulation estimation)
julia> total(:TOTAL_INCOME, :HR, CPHS_income_bootstrap)
102×3 DataFrame
 Row   HR        total         SE
       String    Float64       Float64

   1   HR 1      3.95686e10    6.88132e8
   2   HR 2      6.72443e9     1.96195e8
   3   HR 3      1.93887e10    6.04332e8

 101   HR 95     1.4761e10     4.65155e8
 102   HR 97     1.83399e10    3.58032e8
                         93 rows omitted
```



Plans

Efficient implementations of all the methods in R ‘survey‘. Features for future releases will include:

- Proportion and count estimation
- Variance by Taylor linearization for ‘SurveyDesign‘
- More replicate weighting algorithms (BRR, Jackknife, other types of bootstrap) for ‘ReplicateDesign‘
- Post-stratification, raking, calibration, GREG estimation
- Frequency/contingency table analysis, association tests
- Missing data handling (like R mitools)
- Integration with survival analysis tools
- Integration with GLM.jl
- Out-of-memory integration with SQL databases

Credits

We gratefully acknowledge the JuliaLab at MIT for financial support for this project.

Contributors:

- Ayush Patnaik, XKDR and U Sydney
- Shikhar Mishra XKDR and ANU
- Iulia Dmitru, GSoc and XKDR, U Bucharest, Romania
- Siddhant Chaudhary, Chennai Maths Institute (CMI)
- Sayantika Sengupta, CMI
- Nadia Enhaili, UBC Vancouver
- Fabian Greimel Asso Prof of Economics, U Amsterdam

References

- Model Assisted Survey Sampling (1992)
- Sampling: Design and Analysis, Sharon Lohr, (2010)
- JNK Variance Estimation Literature Review
- R survey package documentation
- Julia Discourse posts here and here
- Unmaterialised attempts samplics/survey.jl and jamanrique/SurveyAnalysis.jl