Czy R (kiedyś) zastąpi SPSS?

Tomasz Żółtak

Instytut Badań Edukacyjnych

29.09.2017 / Konferencja WhyR?

Plan wystąpienia

Badania sondażowe

- W różnych dziedzinach.
  - Nauki społeczne (szeroko pojęte).
  - Badania marketingowe.
  - Badania opinii społecznej (na potrzeby mediów, polityki).
- Ankiety - pytamy ludzi.
  - O to, kim są, kim są (byli) ich rodzice, o poglądy, o wydarzenia z ich życia, jak spędzają czas, co kupują, w jaki sposób pracują, ...
- Relatywnie małe zbiory danych.
  - Typowo kilka-kilkadziesiąt tys. obserwacji (żadno *big data*).
- Prawie wszyscy w Polsce (i nie tylko) używają jednego programu - SPSS.
  - Nie jest on tani.

Specyfika danych sondażowych

- Zmienne *kategorialne*
  - Mało możliwych do przyjęcia wartości (zwykle).
  - Wartościom są przypisane etykiety (sformułowania z ankiety).
    - *zdecydowanie tak*, *raczej tak*, *raczej nie*, *zdecydowanie nie*...
  - Czasem zależy nam, żeby prezentować etykiety.
  - Czasem chcemy coś zrobić z wartościami liczbowymi, do których zostały przypisane.
    - Np. obliczyć średnią... (tyleż nieeleganckie, co popularne).
- Różne rodzaje braków danych:
  - Wartości opisujące wymijające odpowiedzi lub odmowy.
  - Wartości opisujące, że zadanie pytania danej osobie nie miało sensu.

Analiza danych sondażowych

- Bardzo dużo tabel.
  - Różne rodzaje rozkładów: brzegowe, łączne, warunkowe.
  - Czasem chcemy uwzględniać w wynikach odpowiedzi *wymijające*, a czasem nie.
  - Pożądana łatwość przenoszenia tabel do raportów.
  - Testy niezależności chi-kwadrat.
    - Problem *komórek rozkładu o liczebności mniejszej niż 5*.
- *Wielokrotne odpowiedzi*.
  - Kilka zmiennych z odpowiedziami na jedno pytanie ankiety.
- Dużo wykresów kołowych i słupkowych.
- Ważenie obserwacji.
  - Wynikające ze schematu doboru próby lub poststratyfikacji.
  - Ważne uwzględnienie wpływu na oszacowania punktowe, nie na błędy standardowe.

Czego nam trzeba?

- Struktur danych.
  - Oraz narzędzia do importu danych.
- Narzędzia pozwalających wygodnie przekształcać dane *etykietowane* (*label*).
  - A nawet tworzyć w R zbiory ze zmiennymi tego typu.
- Narzędzia do elastycznego tworzenia tabel.
  - Łatwe obsługa odpowiedzi *wymijającymi*.
  - Łatwość zarządzania wyglądem tabeli (np. możliwość automatycznego dodawania wiersza *łącznie*).
  - Obsługa *wielokrotnych odpowiedzi*.
- Narzędzia pozwalających automatycznie używać etykiet (zmiennych, wartości) do anotowania wykresów.
- Narzędzia do łatwego tworzenia raportów.
- GUI, które pozwoli nam to wszystko *wyklikać*?

Struktury danych

- W pakiecie *haven* zaimplementowano możliwość importu danych ze zbiorów SPSS i Stata. W związku z czym zdefiniowano w nim struktury danych umożliwiające przechowanie danych o:
  - etykietach zmiennych,
  - etykietach wartości,
  - *brakach danych użytkownika*.
- Starsze, mniej kompleksowe (i nie całkiem kompatybilne) implementacje w pakietach *foreign* i *Hmisc*.

Operacje na danych etykietowanych

- Pakiety *labelled* i *sjlabelled* umożliwiają:
  - Dostęp do etykiet zmiennych i wartości.
  - Manipulowanie etykietami, w tym dodawanie nowych do zbiorów, w których ich nie było.
  - Konwersję *braków danych użytkownika* na *normalne* wartości/etykiety lub braki danych.
  - Konwersję zmiennych *etykietowanych* na czynniki (*factor*).

Tabele

- Wsparcie głównie tworzenia tabel z parametrami zmiennych ciągłych.
- Brak wsparcia tworzenia rozkładów *wielokrotnych odpowiedzi*.
- Jednak niektóre dostępne rozwiązania są inspirujące.
  - pakiet *tables*:
    - Świetny pakiet do tworzenia tabel o złożonej strukturze.
    - Opis struktury tabeli oparty na formułach.
    - Niestety słaba obsługa braków danych i brak wsparcia dla danych *etykietowanych*.
  - Pakiet *papeR*:
    - Wygodne tworzenie typowych tabel ze statystkami opisowymi zmiennych ciągłych.
    - Brak elastyczności, zwłaszcza w kontekście rozkładów zmiennych kategorialnych.

Ważenie

Pakiet *survey*.

- Posługiwanie się nim jest nieco skomplikowane.
  - Choć istnieją pomysły, jak trochę to uprościć (np. pakiet *srvyr*).
- Stworzony, by realizować dużo bardziej złożone zadania, niż to, czego potrzebują typowi użytkownicy danych sondażowych.

Wielokrotne odpowiedzi

- Na chwilę obecną nie znam pakietów, które wspierały by tworzenie zestawień wyników na tej podstawie.

Wykresy

- Pakiety *papeR* i *sjPlot* wspierają korzystanie z etykiet (zmiennych i wartości).
  - *papeR* w oparciu o funkcje pakietu *graphics*,
  - *sjPlot* w oparciu o funkcje pakietu *ggplot*.
  - Niezbyt duża elastyczność ale łatwe w użyciu.

Raporty

- Ogromne możliwości z Rmarkdown.
- Pewne problemy ze złożonymi tabelami (niezależne od R jako takiego).
  - Typowe formaty docelowe (LaTeX/PDF i HTML) mają tu dużo większe możliwości, niż markdown.
  - Jeśli chcemy produkować złożone tabele, musimy pominąć etap pośredni w postaci reprezentowania ich w postaci kodu markdown.

GUI do klikania

- Chyba jednak najmniej ważne.
  - W praktycznych zastosowaniach i tak dużo operuje się na kodzie (*syntax*) ze względu na reprodukowalność wyników.
  - Choć mogłoby mieć znaczenie dla środowiska akademickiego (i dydaktyki).

Inne pomysły

- Narzędzie do automatycznego łączenia ze sobą wartości zmiennych, aby rozwiązać problem *komórek o liczebności mniejszej niż 5* przy prowadzeniu testów chi-kwadrat.
  - To nie jest działanie zbyt eleganckie... ale wiele osób byłoby tym bardzo zainteresowanych.
  - Takie łączenie musi być robione *rozsądnie*.
  - Wskazane byłoby dołożenie pewnej logiki mówiącej o tym, które odpowiedzi można za sobą łączyć, a których nie (albo które lepiej ze sobą łączyć, niż inne).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

- Mamy podstawy do analizy w R danych sondażowych.
  - Struktury danych - w ramach *hadleyverse*.
  - Podstawowe narzędzia do przekształcania danych *etykietowanych*.
- Dużą zaletą R są raporty i możliwości wizualizacji.
- Deficyty na poziomie możliwości prowadzenia (skądinąd bardzo prostych) analiz.
  - Potrzebny *kombajn* do tworzenia tabel z rozkładami zmiennych kategorialnych.
  - Potrzeba możliwości *łatwego* uwzględnienia w analizach ważenia.
- Najpoważniejsza przeszkoda to *czynnik ludzki*?.
  - Być może łatwiej przekonać do korzystania z R firmy badawcze, niż naukowców (uczących studentów).

Podsumowanie

Dziękuję za uwagę!

Tomasz Żółtak
tomek@zozlak.org