Now That's Fresh Water*

Trends in India from the 1992 Demographic and Health Surve

Bilal Haq and Ritvik Puri

01 April 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

Contents

*Code and data are available at: https://github.com/haqbilal/India-1992-DHS-State-Findings

1 Introduction

The period 1980-2000 featured data science at the forefront of a developing era of statistical modelling, and several countries were finally jumping on the bandwagon. Acknowledging that a large scale survey on the population of India had never been done before, K.B. Pathak was ready to forge the path forward on behalf of his country. He was the director of the International Institute for Population Sciences (IIPS) in Bombay, and his team headed the Project to Strengthen the Research Capabilities of the Population Research Centres in India in 1991. The primary objective of this campaign was to finally provide state-level and national-level estimates of fertility, infant/child mortality, family planning practices, maternal/child health care, and the utilization of services provided for mothers and children. The National Family Health Survey (NFHS) took on this mission, and as a byproduct, provided high quality data to statisticians for analytical research on a range of population and health studies.

The release of this report came with a call from the IIPS for further analyses of the NFHS data, among researches both in and outside of India. The final report was made publicly available to anyone in the word through the United States Agency for International Development's (USAID) Demographic and Health Surveys (DHS) program. The report contains a thorough record of household/individual respondent characteristics. Namely, data on marriage, fertility, family planning, utilization of antenatal services, vaccination, child feeding practices, nutritional status of children, and knowledge of AIDS. Alongside, interstate variations on key indicators were also included, allowing for regional comparisons to be made. As independent researchers and data scientists, we have taken on the IIPS's call to further analyze this data, and report our findings and comments in this paper.

The report onwards features three sections. Firstly, we look at the dataset provided on pages 31-32 of the NFHS 1992 Final Report. Plots and directed acyclic graphs are used to visualize the data and relationship between variables. Moving on, we share our findings and the implications of results and the effect they might have on public policy in India. We also align our own findings with those of the NFHS, and analyze the progress made over the last 20 years, demonstrating the effect that data analysis has had on the development of the country. Finally, we discuss the methodology used by the NFHS and describe how they were effective in achieving their goal. We also propose some critiques that can aid in the future of data analysis in India and other third world countries following in their footsteps. Our report is carried out using R (R Core Team 2020), alongside various libraries cited (see References). R markdown (**citeRMD?**) was used for compilation and presentation.

2 Data

2.1 The Dataset and Variables

The NFHS's final report from 1992 is publicly hosted on the DHS program's website in PDF format. The data has been sourced in a reproducible way, with a script for downloading and obtaining the relevant datasets. It is an in-depth review of statistics such as access to clean water and amenities, fertility, family planning, child care, and health, on a sample of women and children, with the goal being to summarize the status of women and children, as key factors towards growth of the country, while identifying any present flaws, in hopes of administering aid where possible. In terms of privacy, no underlying information has been kept that could point towards or reveal any individual that took part in the survey. Note that a constraint of this dataset is that it cannot be truly compared to similar reports from other organizations, because of differing methods in categorization as well as geographic stratification. Time constraints may also affect results from different surveys.

Our focus is on the state findings table on pages 31-32 of the report. The full table is shown on the next page [SHOW BOTH TABLES ON NEXT PAGE]. In particular, we will look at the columns: percent of illiterate females, households with drinking water, mothers receiving antenatal care (that is, special care while pregnant), births delivered in a health facility, and fully immunized children. These variables were selected because they paint the picture of health conditions for those living in India, and the status of pertinent

features such as access to water, and amenities for pregnant women and their children. This information is given for each of the 25 states, as well as in total for all of India. Information from some columns was unavailable or omitted because it involved 25 or less people, in which case they were left out of analysis. A table for the first 5 states of the dataset with filtered variables is shown below [TABLE HERE]. All quantities are percentages of the sample for that state, unless otherwise specified.

2.2 Methodology

In formulating the survey questionnaires (see Appendix B), a Questionnaire Design Workshop was held in Pune, September 1991. There, representatives from IIPS and other organizations such as USAID, designed the contents, which were designed with India's low contraceptive prevalence in mind. Modifications were made to the questionnaire that kept in consideration the Indian social and cultural environment as well as the NFHS objectives. Individual states also included state-specific questions pertaining to issues of importance in those regions. Three questionnaires were created; the Household Questionnaire, Woman's Questionnaire, and Village Questionnaire.

The Household Questionnaire listed counts of residents in each sample household, and visitors who stayed from the night before. Information collected included age, sex, marital status, education, occupation, and household conditions such as access to water and toilets. It's main purpose was to identify women eligible to participate in the Woman's Questionnaire (women in between 13-49 years of age). The Woman's Questionnaire then gathered information on the respondents' background, reproduction (i.e. births given), contraception, health of children, fertility preferences (i.e. desire for children), and nutritional status of children (i.e. height and weight). Lastly, the Village Questionnaire collected data on the amenities accessible to villages, such as electricity, water, transport, and educational and health facilities.

A three-stage sample design was used in each state: first selecting cities, then blocks, and finally households. These selections were based on enumerations of the cities and blocks from India's 1991 census. The cities were divided into three strata: (1) self-selecting cities (i.e. volunteers), (2) district headquarters, (3) all other cities. In each strata, selection of the required number of blocks was followed by selection of an average of 20 households. This accounts for possible bias in population densities of different cities, as well as geographical/regional bias from different locations.

The methodology was prone to one major weakness, accessibility of the survey. Data collection teams were noted to struggle in reaching sampling units located in hilly regions, due to lack of roads and other transportation. For example, in the state Uttar Pradesh, teams covered the sampling units on foot. Further, security was another concern. During December 1992, communal riots caused the presence of bandits in some villages. In this case, sampling units in Madhya Pradesh were avoided, and the data collection could not be fully completed. Moreover, unseasonal rains that were not forecast delayed data collection in Karnataka, Kerala, and other Northeastern states. Lastly, teams failed to establish contact with some households, and as many as possible were attempted a second time at a later date to minimize the nonresponse rate, however in some cases it was unavoidable.

2.3 Visualizations

For all comparisons made in this section, we look at each state as a different case, i.e. an example of the conditions met in that state. Thus, we can compare between states, but not within a state, because of the nature of the data.

The first comparison we make is between the percentage of illiterate females and the percentage of vaccinated children. This will tell us the effect education of the mother has on the benefit of their child. The graph below [GRAPH HERE] shows that as illiteracy rate goes up, the rate of vaccination goes down. This seems indicative that educated women are more likely to vaccinate their children.

Furthermore, we also want to check the relation between percentage of mothers receiving antenatal care and percentage of births delivered in a health facility. After plotting the two variables together and fitting a

linear model [GRAPH HERE], we see that states with a high rate of antenatal care also have a high rate of births in a health facility. From this we can interpret that states with a priority on health care take better care of their pregnant women and children. Thus it seems likely, and is indeed the case as demonstrated by [GRAPH HERE], that states with a high illiteracy rate also have a low rate of births in a health facility.

Another type of relation we want to study is the interaction between our variables. That is, we want to know if a change in one is affecting a change in another, on the yield of a third. To aid in this process, we define our two yields, or outcomes, as the immunization rate and birth in health facility rate of children. This makes sense as these two variables serve as indicators of the wellbeing of children in the country. Then our 3 predictors are the illiteracy rate, access to water rate, and antenatal care rate.

Moving forward, we will define the 5 rates above as being high, if they are greater than the grand mean (corresponding to the row for India), and low if they are less. We will create 5 new indicator variables for each rate, taking the value 1 if the rate is high, and -1 if the rate is low. This will allow us to treat the data as coming from a 2^k factorial design, where $k = 3$ is the number of levels, with 2 outcomes. The interaction plots for each of the $k! = 3! = 6$ combinations below [PLOTS HERE] show that we have interactions between the following of variables, as represented by the directed acyclic graph [DAG HERE], where a variable points to a variable that it influences, treating the outcome variables as influenced by the predictor variables. Note that the outcomes also appear to have an affect on each other, where as shown by the graph [GRAPH HERE], as states with a higher birth in a health facility rate, also have a high vaccination rate.

2.4 Summaries

The overall findings from the survey are summarized in this section, as they pertain to our variables of interest. 57% of all females age 6 and above are illiterate, with about 9% having a high school education or more. 68% of households have access to some form of drinking water, from a pump or pipe. We have that 62% of mothers received special care while they were pregnant, but only 26% of mothers gave birth in medical facilities. Finally, only 35% of children were immunized. Using these findings, we can see that key areas for improvement in India's health sector are availability of medical facilities, as we know from above that those receiving antenatal care would give birth in those facilities if they were accessible. Further, the low vaccination rate can be combatted by spreading awareness of the importance of vaccines, outside of schools, because the majority of females could not attend school to learn about them. A core, systemic, issue in India is the high poverty rate, which likely is responsible for many of the above issues alone. However, the NFHS survey failed to account for this variable. If they had recorded the economic status of their sampling units, another form of bias could be eliminated, and the results purified. Thus, in the future, accounting for this source of variation (both within and between states) can lead to more reliable results.

3 Results

Requirements: - Summary statistics - Tables, graphs, images, maps - Statistical analysis of the above - Associated text with all figures - Plot results where possible and talk about them - Strictly relay results (regression tables must not contain stars)

4 Discussion

Requirements (each is a half-page subsection): - What is done in this paper? - What is something that we learn about the world? - What is another thing that we learn about the world? - What are some strengths and weaknesses of what was done? - What is left to learn or how should we proceed in the future?

Appendix

A Additional details

References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.