PhD in Britain, student records 1917-1959, an anonymised sample database

Based on Renate Simpson's 'The Development of the PhD degree in Britain, 1917-1959 and since: an evolutionary and statistical history in Higher Education', published 2009. The author and book are referred to as RSS and DPhD respectively in this documentation.

Summary

DPhD reviewed the development of the PhD degree in Britain, drawing on archive materials including student records from seven Universities which included nearly half of all doctoral candidates during the period. These Universities are Cambridge, Edinburgh, Imperial College, London School of Economics, Manchester, Oxford, and University College London. Analysis of the records informed DPhD's Part 2, 'The British PhD in Numbers'. Details were collected from a sample of students, as described in pp220-233. All page and Table references in this documentation are to DPhD unless specified otherwise.

The variables in the database refer to the student's previous University, whether UK or overseas; department and faculty where registered; outcome of studies and length of time taken until successful or not continuing registration. For some universities, age, sex, staff status are included. A full data dictionary is included

The revised database is accompanied by a data dictionary, which lists each variable on the database, describes the values that the variable can take, and provides the number of students in each University recorded with each of those categories.

Permission to use the database is conditional on neither attempting to identify individuals nor appearing to divulge individual identities of students.

Development of the database

The work for DPhD was begun in earnest in the 1980s and much of the analysis was completed in the 1990s, before current standards of computer hardware and software had developed. The details of students from the University administrative archives were recorded on cards. They were transferred to a computer database of student records which was structured and analysed within the free software EPI-INFO, used widely in developing countries for health statistics. Various revisions and additions to the database were made by RSS. The version that survives was created on 2nd April 1997 and matches the tables in DPhD in all but minor details. That version is referred to as the 'original database'. The record cards and earlier versions have not survived.

The revised database which accompanies this technical documentation contains all the information in the original database, except the student names.

Some variables have been derived from the original database so that Tables in DPhD can be more easily reproduced and extended, such as those referring to full-time students, to staff, or to overseas students from countries of the Commonwealth or Empire.

Some information is included which was not used in DPhD. This is mostly information which was only available for a minority of the seven universities studied, or was not comparable between Universities,

for example on the retrospective registration for doctoral studies dependent on completion of other qualifications. Four variables at the end of the dictionary contain this extra information.

The descriptions of each variable and its categories draw on DPhD and on hand-written notes by RSS in a single small clip file. This and the original database are kept in family archives.

The notes on the following pages describe the sampling scheme is reflected in the database, and checks that were made to assess the consistency of the database with information in DPhD together with the minor discrepancies found.

The DPhD student sample

DPhD Table I-2 on p223 defines the population of 23,510 which is represented by a sample of 9,606 records. The population is a multiplication of the sample by the sample weights, which have been added to the original database. The database is a weighted version of the sample, and "It is this latter weighted figure [23,510] which forms the basis of the analyses presented" (p225).

The sampling fractions were constant within a decade (of students' admission date) in each University, as given in Table I-2. The sampling fractions are, with three exceptions, singular fractions: 1 in 2, 1 in 3 and so on. The sample weight is a whole number: 2, 3 and so on. In the database, the weighting in all these cases was achieved by replicating the records by the inverse of the sampling fraction - two identical records for a sample of 1 in 2, and so on.

The exceptions are Cambridge 1950s (sampling fraction 13.9%), Manchester 1950s (28.8%) and UCL 1940s (66.7%) (all given in table I-2 and discussed in nearby text).

> The Cambridge 1950s sampling is described on page 226 as one in 9 plus one in 32 of the remainder, resulting in a sampling fraction of 1/7.2. The sample of 380 have unique names. 304 (80%) names appear 7 times, and the rest 8 times, to give the overall weight of 7.2.
>
> The Manchester 1950s sample was limited for Arts and Social Sciences due to lack of data for 1955-59 (pp225-6). The sample of 390 (Table I-2) have unique name-Faculty combinations. The sampling fraction was 1 in 2 for Arts and Social Sciences, and the duplicates were given a year of admission five years later, to simulate the population of 1955-59. The sampling fraction was 1 in 4 for the much more numerous Science and Technology students. Six student names of Technology appear only twice, and eight student names in Social Science appear four times. Notes suggest that their classification into Faculties was changed after the sample was chosen, to overcome the different classification of subjects to Faculties by different Universities. Faculty numbers are consistent with DPhD.)
>
> The UCL 1940s sample was 2 out of 3. Half of the names appear twice and the other half are not replicated.

The sample weight for a student in the sample is reflected in the database by including replicates as described. The weight has been included on the revised database, and a sample ID unique to the student has been given instead of the student's name.

Some analysts may prefer to make a database with only one record per sampled students, and use their analytical tools to weight each sample member. This can be done but four anomalies would have

to be dealt with as follows, which appear to be errors of some sort in the duplication of records. In all other cases, the records for the same student are identical:

> SampleID 6302 has two records (sample weight 2). Variable 'PrevUnii' (previous university) has A (Australia) and NZ (New Zealand) on the two records.
>
> SampleID 7532 has seven records (sample weight 7). Variable 'Duration' (duration of studies) is 42 months on all but one of them, on which it is missing. Variable 'DurPrev' (Duration of research prior to registration for Doctorate) is 42 on two of them but missing on the other five.
>
> Sample ID 2956 has four records(sample weight 4). Variable 'DurPrev' is 5 on all but one of them, on which it is missing.
>
> Sample ID 8551 has four records(sample weight 4). Variable 'DurPrev' is 6 on all but one of them, on which it is missing.

Some statistical work on the representativeness of the sample was carried out in 1992. A possible bias towards slightly more social sciences in the sample (by one % point overall, and a maximum of 2.5 % points in any University or decade) was identified, but could be accounted for by the different classifications in the population figures. Paper copies of this analysis survive.

Consistency of the database with DPhD

The database has the same number of sample members, with the same weights to represent the population of students at each University in each decade, as reported in Table I-2.

The database was used to successfully reproduce at least one table from each chapter of part 2 of DPhD. The only exception was the chapter on age, where one student aged 24-29 in the database must be aged 19-23 in DPhD, for the two to be made consistent. The student must be a female Arts student from a Home University, but the specific student could not be identified.

Age at completion of studies is not a variable on the database, but is used in Chapter VIIB. It is recommended to be calculated as (age +0.5 +(duration of studies in months)/12). This reproduces Table VIIB-5 closely for mean and median (exactly or one year out for 3 of 12 comparisons). It has differences from the DPhD distribution between age groups that are not biased towards older or younger groups (25-29 is higher in DPhD but lower in 21-24 *and* in older age groups). Notes suggest that RSS recorded month and year of birth, and computed age and age at award from this; month and year of birth are not on the surviving computer records, and this would account for the slight inconsistencies in the recommended calculation of age at completion of studies, compared to DPhD. The recommendation replicates Table VIIB-5 much more closely than the inaccurate use of age as if it was exact, ie without adding 0.5.

Ludi Simpson, 20th February 2015 ludi.simpson@manchester.ac.uk