HarvardX Research: worldmap of gender composition data specification

The interactive visualization appearing at [URL] and at [URL] is based on data on HarvardX enrollment for courses offered through the edX platform. Enrollment is defined as individuals on HarvardX course lists. This document aims to describe (1) the way data was prepared and its possible sources of error, (2) questions the visualization helps answer, (3) possible misinterpretations of the data.

Preparation and possible sources of error

The data was prepared by parsing self-reported gender information provided at registration by current HarvardX students, and combining it with geographic location data as inferred from IP geolocation or mailing address parsing as of November 17, 2013 (or the date specified on the visualization). Proportions of males and females when Undisclosed and Missing categories are ignored are inferred under the assumption that any missing data is Missing At Random for ignored data. The Python code for extracting the data can be found at nesterko.com/visuals/worldmap-gender_2/data_for_hx_gender_worldmap.py. Resulting dataset is at nesterko.com/visuals/worldmap-gender_2/data_for_hx_gender_worldmap.csv.

Specification of possible sources of error:
1. Our data stores may have errors recording student registration information (on the scale of 50 students per course).
2. Inferred proportions of males and females are based on a statistical method and carry inherent uncertainty. For a sample size of less than 100, the margin of error can exceed 5%. Error levels need to be evaluated further if the Missing At Random assumption is not satisfied by missing data.
3. Baseline numbers of students from each country used for estimation are subject to geolocation error. The IP geolocation database we use may have imprecisions, and the address parser, which is used as a supplementary location information source, carries errors. In our analysis, address parser does not match IP geolocation about 8% of the time when both are available for an individual.

Possible misinterpretations

1. It is important to note that this figure represents all registered students, without any kind of classification filter based on activity in the course. It may be that the gender composition of active students is different than the gender composition of registered students.
2. As noted above, inferred proportions of males and females carry estimation error. We recommend supplementing analyses basing arguments on exact percentage estimates to incorporate measures of uncertainty in these estimates, such as standard errors or confidence intervals.
3. As noted above, we cannot verify the accuracy of the location data, and we know there are

several possible sources of error. For analyses where precise residence data is important, it might be better to include other location methods such as course registration surveys and other methods.