HarvardX Research: gender composition data specification

The interactive visualization appearing at http://harvardx.harvard.edu/harvardx-insights/gender-composition 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) possible misinterpretations of the data.

Preparation and possible sources of error

The data was prepared by aggregating self-reported gender counts of HarvardX registrants as of the date displayed on the visualization. The Python code used to extract the numbers can be found at http://ow.ly/tLa1A. Resulting datasets are at http://ow.ly/tLa88 and http://ow.ly/tLabv.

Specification of possible sources of error:
1. Gender composition is inferred based on self-reported gender at registration, and thus may not reflect the true gender proportions in case there is bias in how registrants report gender.
2. The 'Only male/female' option excludes missing gender information and self-reported gender of 'other' from the charts. The resulting calculated percentages are prone to the same errors as described in 1.

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 or certified enrollees is different than the gender composition of registrants.
2. As noted above, we cannot account for the possible systemic biases in self-reported gender, given the wide range of geographic locations and cultures registrants represent. For analyses where precise gender data is important, it may be appropriate to further reconcile possible biases induced by various factors such as cultural environment, language proficiency, and others.