HarvardX Research: age composition data specification

The interactive visualization appearing at harvardx.harvard.edu/harvardx-insights/age-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 calculating self-reported age at registration on the edX platform, for all HarvardX registrants as of the date displayed on the visualization. If the resulting age is calculated as less than 6 years or greater than 100 years, the data point is marked as 'missing'. Usually, there are less than 1,000 such occurrences across all of HarvardX registration lists. The Python code used to extract the numbers can be found at http://ow.ly/tKZfJ. Resulting datasets are at http://ow.ly/tKZDo and http://ow.ly/tKZPp.

Specification of possible sources of error:
1. Age composition is inferred based on self-reported year of birth at registration, and thus may not reflect the true age proportions in case there is bias in how registrants report gender.
2. The resulting calculated percentages after checking the 'Normalize' option 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 age composition of active or certified enrollees is different than the age composition of registrants.
2. As noted above, we cannot account for the possible systemic biases in self-reported year of birth, given the wide range of geographic locations and cultures registrants represent. For analyses where precise age data is important, it may be appropriate to further reconcile possible biases induced by various factors such as cultural environment, language proficiency, and others.