Alan Ridel

We know that there were a large number (25,000 books published during the victorian era) of books, we have a lot of information about gender and year level stats.

no corpus that exists reflects the population of published novels during this period effectively.

The Chadwyck-Healey corpus is particularly bad, 50% of the data comes from male authors published before 1876 even though this was only 15% of the population.

Random sampling of the population is not really possible because we don't actually have a complete database of all novels published during the victorian era.

instead we do quota sampling.

We divide up the population into categories based on year and gender and manually encode a randomly selected chapter.

- Not a representative sample
  - overrepresents authors who wrote more than one novel
  - over represents novels published in multiple volumes

Maybe there's a bias in which things were published or which types of genres tend to do multi volume things

The solution is to use post-stratification as a way to do analysis of granular distinctions after the fact:

- e.g. novels published by women in 1940
- novels involving trains

Revision #2

Created Wed, Nov 20, 2019 7:41 PM by kenneth

Updated Thu, Nov 21, 2019 12:58 AM by kenneth