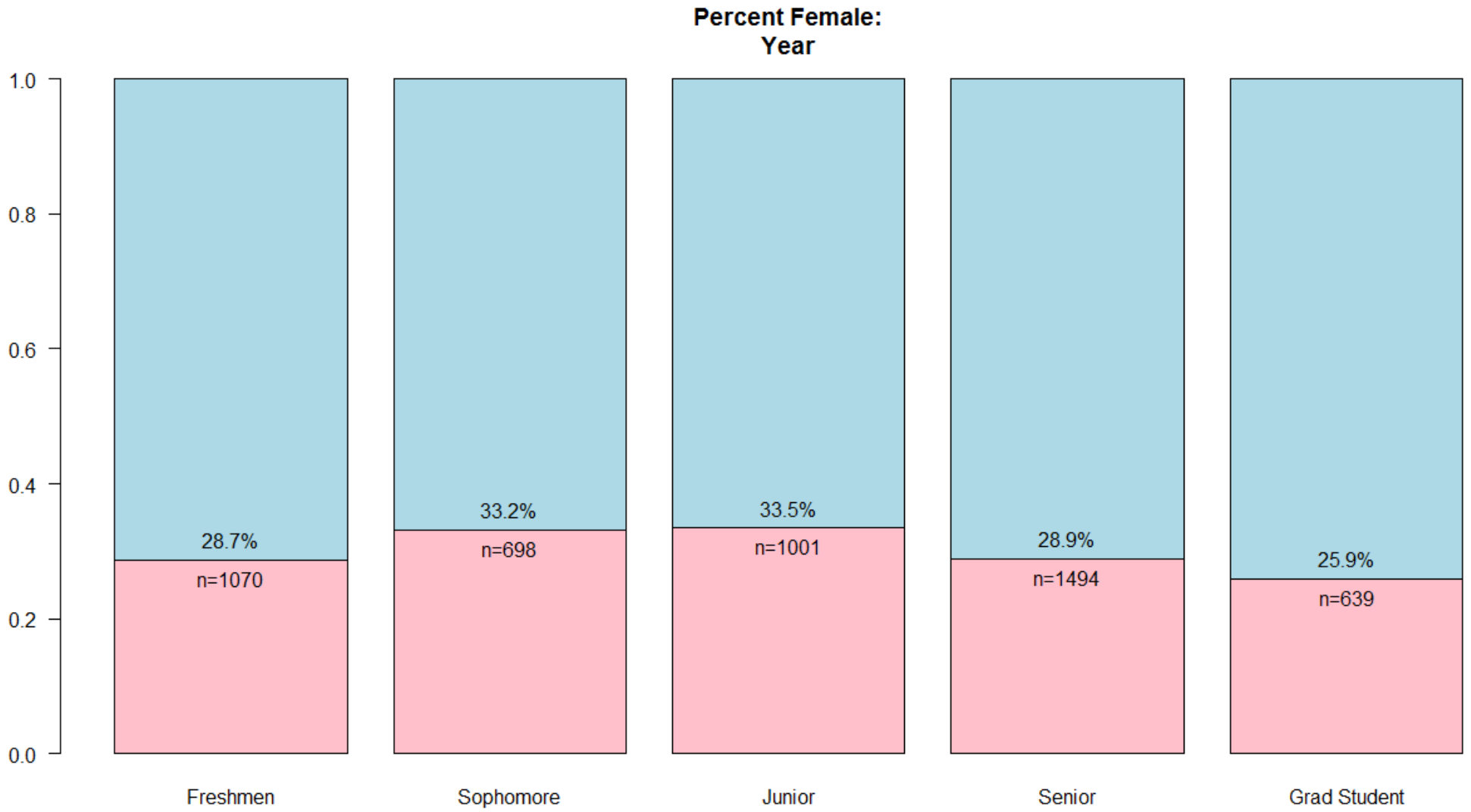
RPI Gender Breakdown Estimate

https://github.com/stensonowen/RPI_ratio



Overview

1) Python Scraping
2) Statistics
3) Plotting in R

Python Scraping

- Data courtesy of Rensselaer Directories
- Python's Requests library used for download
  - Session object necessary to follow redirects
- Takes about 15 minutes
- Gathers about 10,000 entries
- Relevant functions: get_by_index(n), parse(html), extract(data), fetch(n)
- Demo

Statistics

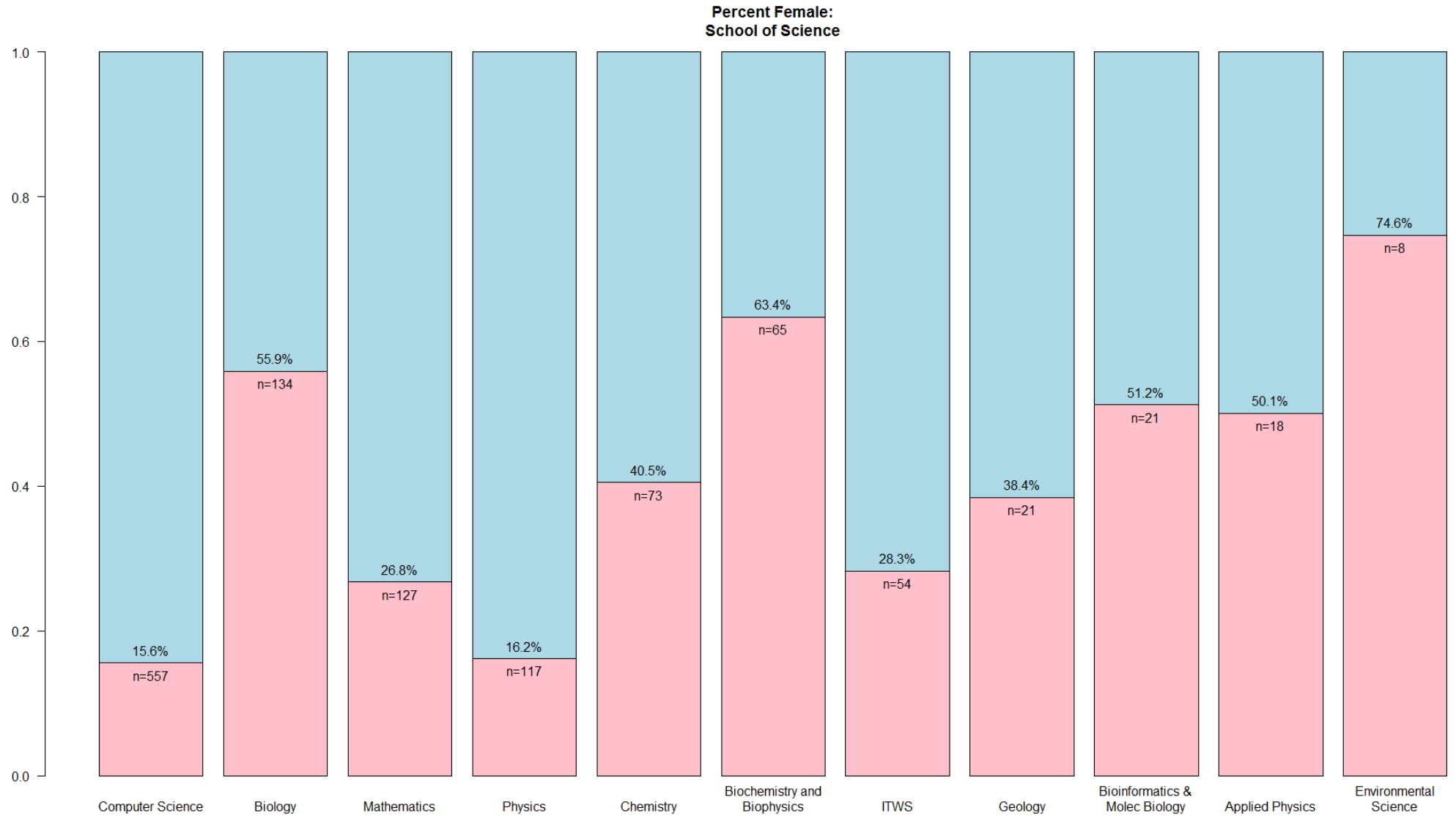
- Name data from ssa.gov/oact/babynames
- Data in the form “`<Name>,<[M|F]>,<Count>`”
- Data from '93-'97 gives 37k names, 18M people
- `Count[i]` and `sum(Count)` give $P(Name)$
- “`<Name>,M,<CountM>`” and “`<Name>,F,<CountF>`” give $P(Name \mid Gender)$
- Bayes' Theorem gives

$$P(Gender \mid Name) = \frac{P(Name \mid Gender) * P(Gender)}{P(Name)}$$

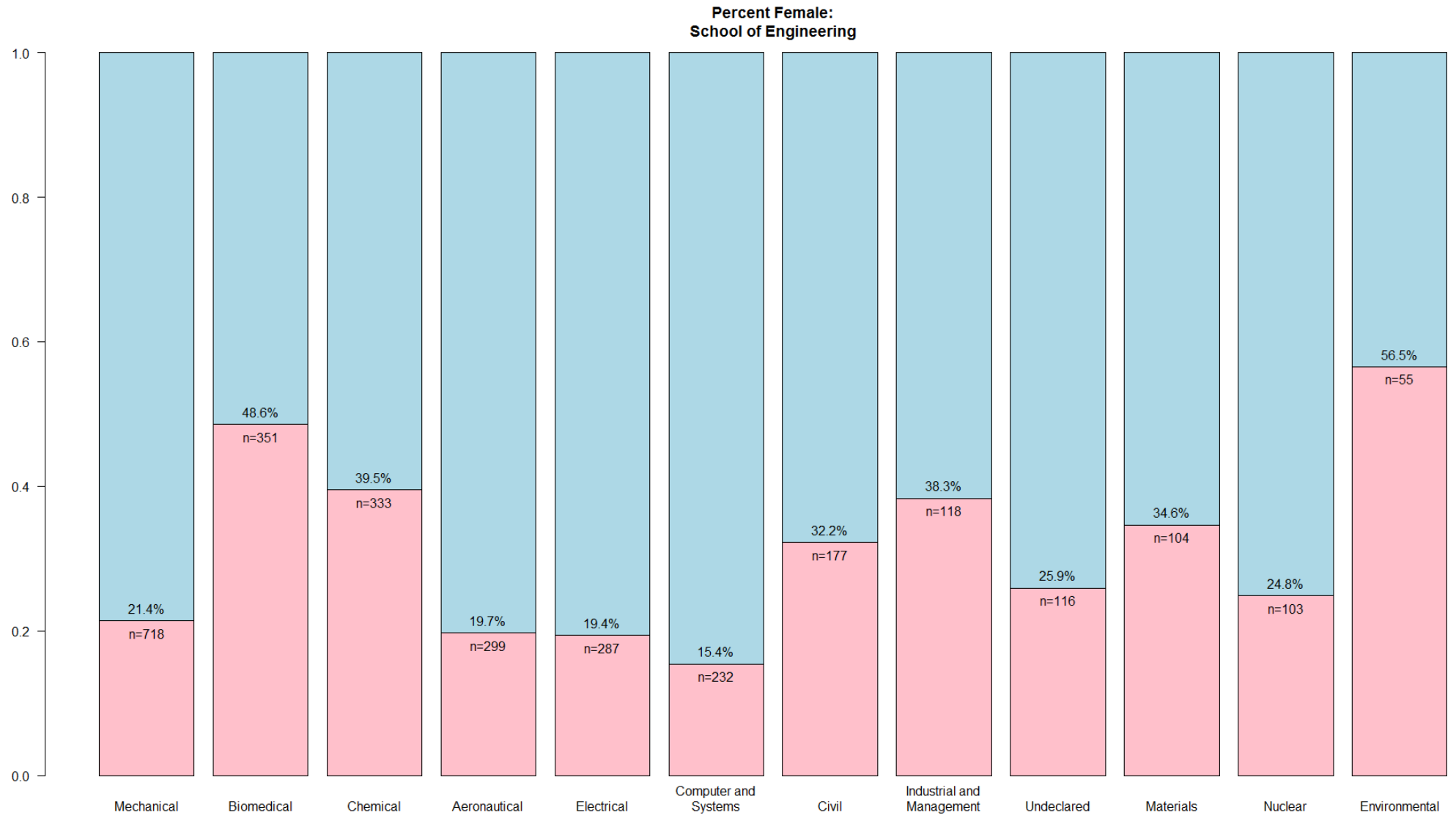
Plotting

- Generic `barplot()` in R
- Function `annotated_barplot()` automatically plots subsets of the results, and formatting can be optionally specified

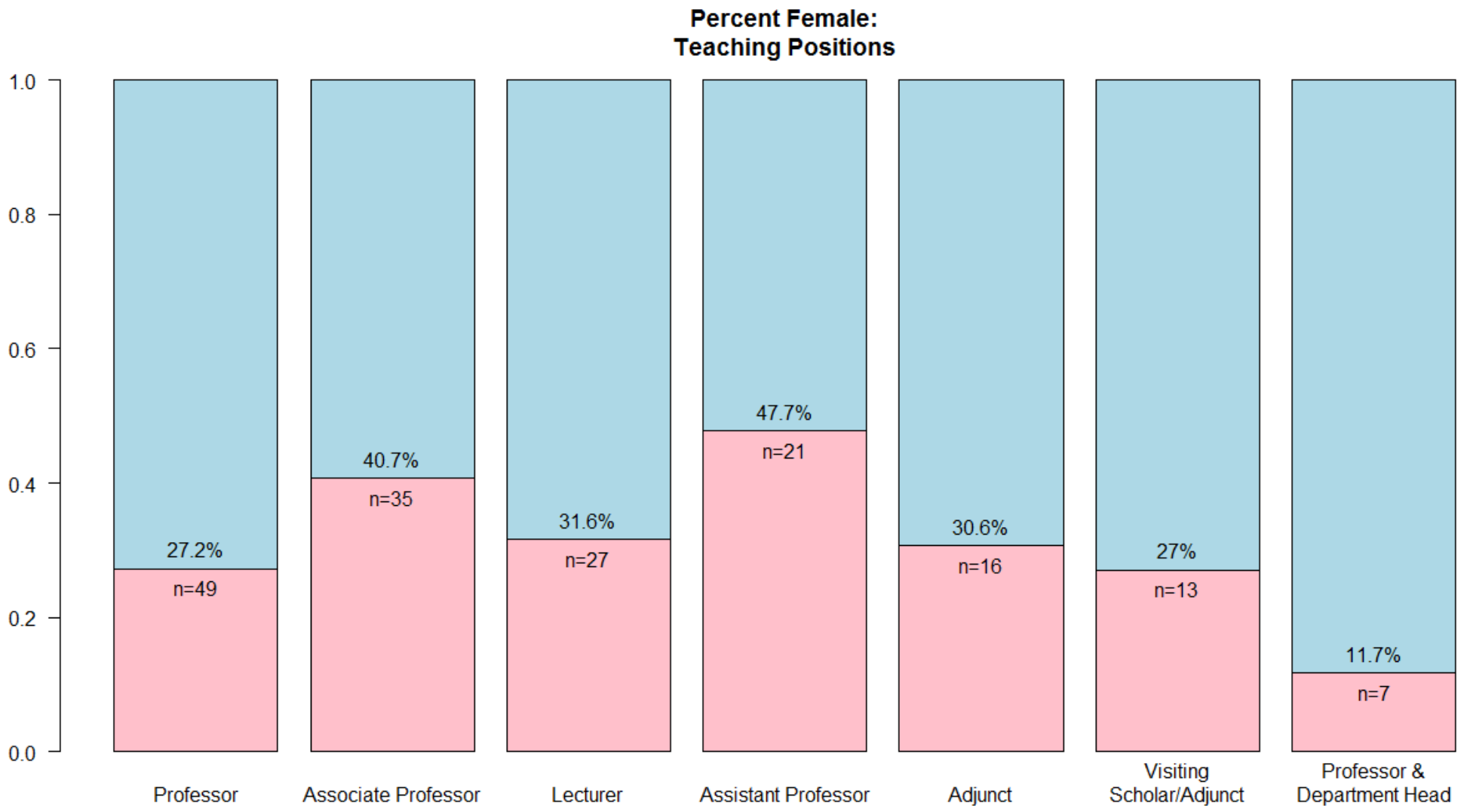
Example: School of Science



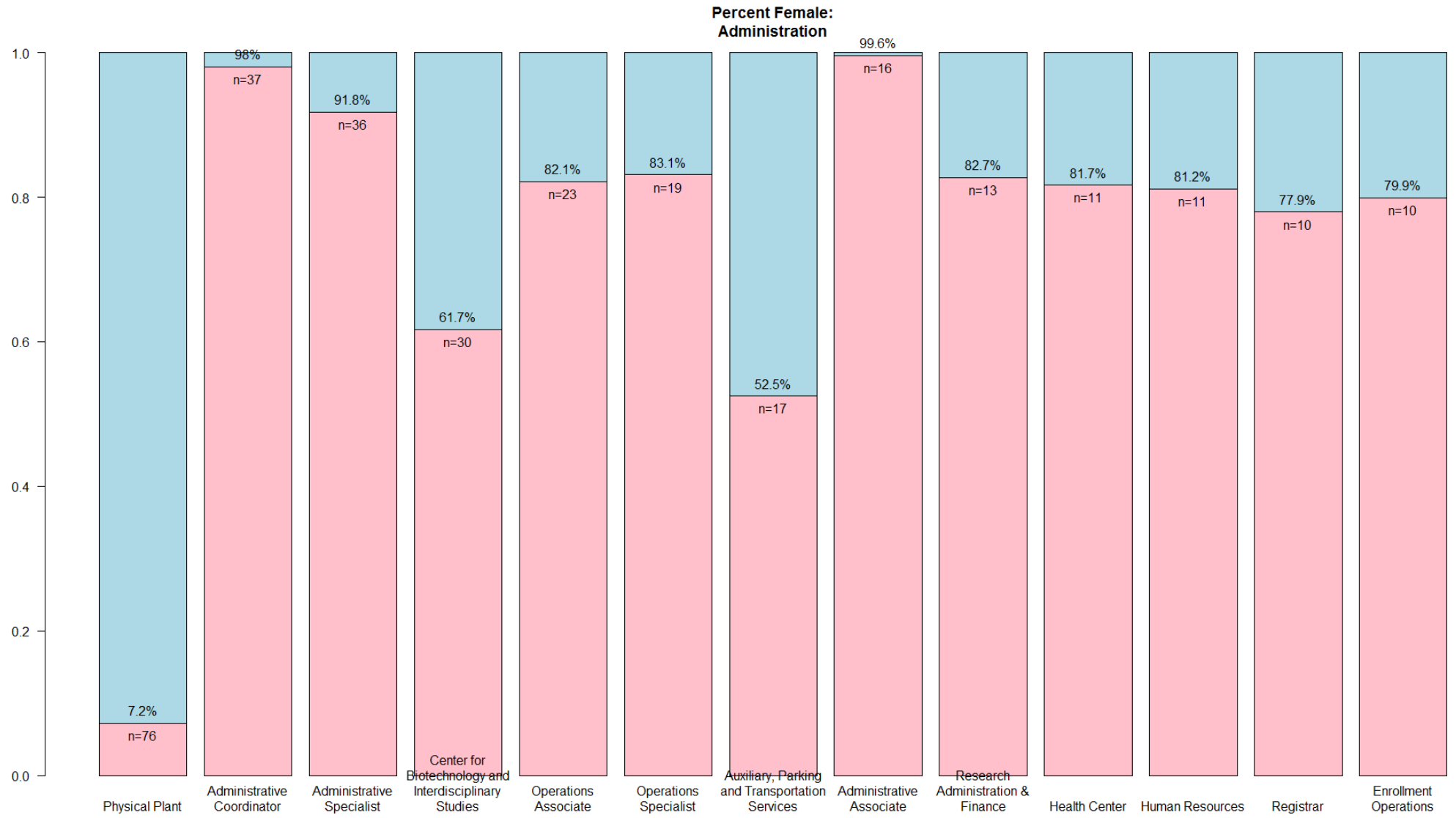
Example: School of Engineering



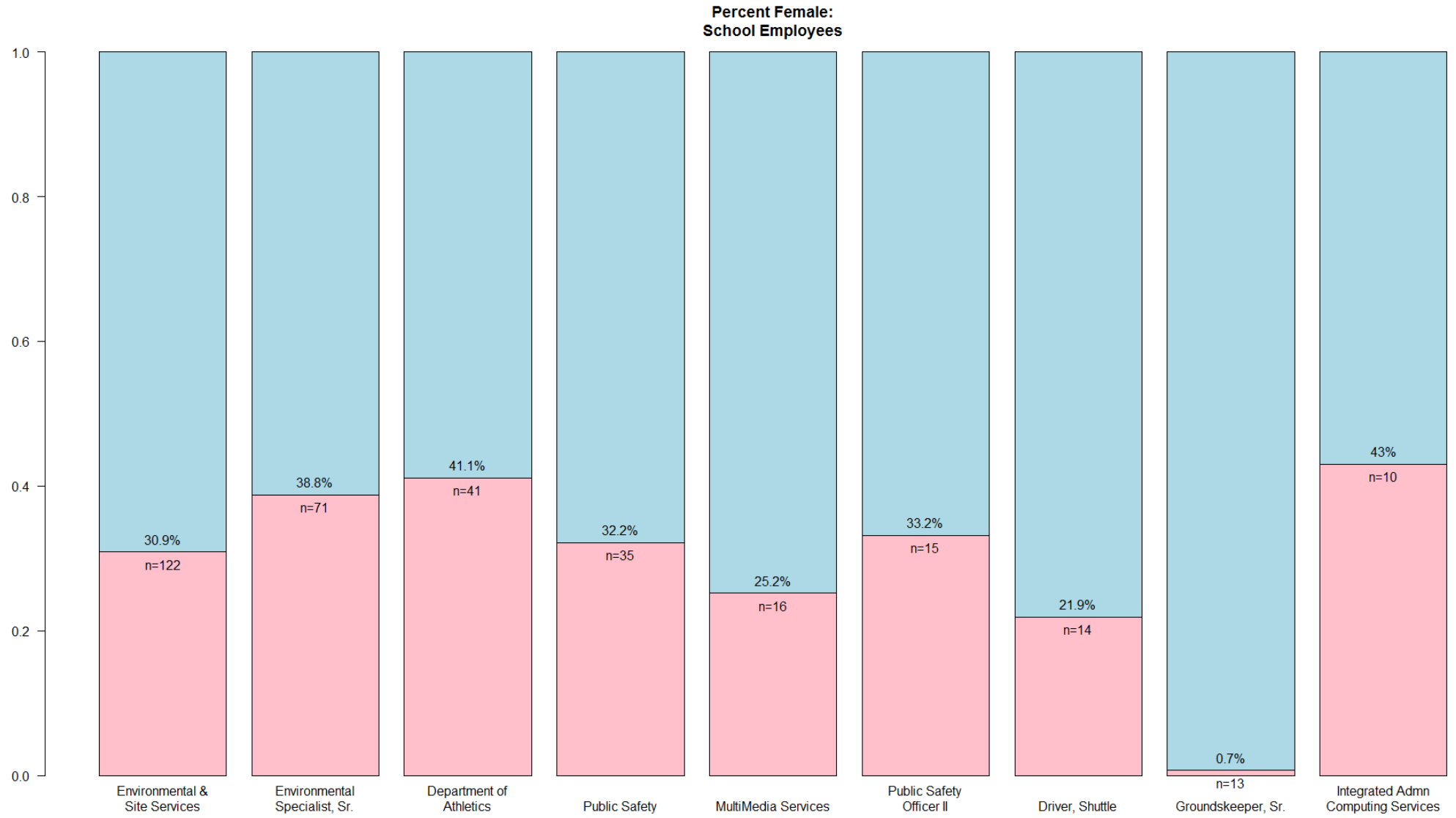
Example: Teaching Positions



Example: Administration



Example: Employees



For More:

- Slides and more images at https://github.com/stensonowen/RPI_ratio
- More data in results.csv file
- More student data at https://rpidirectory.appspot.com/