Chapter 4

The Legacy of Eugenics in Statistics

As identified in Chapter 3, the foundation of regression modelling in statistics was eugenic. More broadly, historians of science and mathematics have identified the close contact of the eugenics and statistics movement across their foundational years, circa 1890-1930. During these years, Galton, Pearson, and Ronald Aylmer (R.A.) Fisher single-handedly[1] the framework of modern statistics: Galton invented regression; Pearson, correlation and the χ^2 test; and Fisher, ANOVA, hypothesis testing, likelihood, and consistency, among others [49, 43]. Another link between these men was their eugenic beliefs, which each professed publicly and wrote about extensively [53].

The question of interaction between statistics and eugenics in the work of these men has been a scholarly focus in the history of statistics for several decades [2]. However, little research[3] has been done on the trail of eugenic thought in the growing field of statistics during and after these fundamental years. Since Galton and Pearson's fundamental work on statistics was linked to eugenics, t is possible that eugenic ideas were linked to statistical pedagogy throughout the 20th century. To address this gap in

[1] According to the most common canon—as explained in the introductory paragraph of Chapter 1, many concepts across the history of science are not named for their true 'inventor,' being the first to discover or develop it, but rather for a later 'inventor' with better publicity. Furthermore, to ascribe full responsibility to these three men would be incorrect, as both Pearson and Fisher had large, well-funded laboratories dedicated to statistics, while Galton before them also enlisted others to gather his regression data and to help with mathematical proofs and computations. However, for the purposes of this chapter, divergence from canonical inventors is impertinent, as will soon be made clear.

[2] A non-exhaustive list of works focusing on this interaction includes the following, listed in chronological order by date of publication [24, 13, 64, 55, 69, 57, 58, 53, 15, 43, 8, 18].

[3] Some examples of research in this vein are Francisco Louçã's "Emancipation through Interaction: How Eugenics and Statistics Converged and Diverged" [53] and Barbara Arneil's "The Intersection of Ableism, Domestic Colonialism and Statistics in Britain from Bentham to Galton" [5].

knowledge, I gathered text data from statistics textbooks published between 1880 and 1970 and performed text-mining analyses to understand the extent to which eugenics appeared alongside statistics in pedagogical material.

4.1 Previous Work

This study draws inspiration from two previous works: "The foundations of statistical science: A history of textbook presentations" by Alan Agresti and *Inheriting Shame: The Story of Race and Eugenics in America* by Steven Selden [1, 85].

Agresti's "The foundations of statistical science" presents a close qualitative analysis of statistics textbooks at the inception of the mathematical statistics movement (ca. 1911)[4] through the 20th century. Beginning with Arthur Bowley's *Elements of Statistics* (1901) as a 'control group,' Agresti tracks the changes in the pedagogical materials available for training statisticians as a representation of the progression of the discipline as a whole. The large majority of his focus is on books explaining "statistical *theory* or what later became identified as *mathematical statistics*" though he includes "*Statistical Methods for Research Workers* by R. A. Fisher (1925) and the more reader-friendly *Statistical Methods* by George Snedecor (1937)" due to their disciplinary impact and popularity [1, p. 658]. His findings may be summarized by the simple statement that, in English-language works[5], the "key modern concepts of statistical science" [1, p. 657] — correlation, regression, chi-squared tests, t-tests, analysis of variance, method of moments parameter estimation, maximum likelihood estimation, and principal components analysis — were all developed and widely circulated within the first half of the 20th century. Agresti's reads like a literature review of the textbooks: describing how they were organized, which concepts were included, and counting the number of pages dedicated to each topic. He includes 47 different textbooks, not including new editions, with publication dates ranging from 1901 to 1995. However, only 12 receive full treatment, including contextualization of contemporary response via reviews published in academic journals. The earlier textbooks tend to have a fuller treatment, due to the exponential increase in the number of publications post-World War II. It is unclear

[4] Agresti claims that this year would be the start of the mathematical statistics movement since it is the year that Pearson established the Department of Statistics at University College London (the first in the world) and Yule, his student, published the textbook *Introduction to the Theory of Statistics*. I object to placing the start date so late, as many important mathematical contributions to statistics occurred in the late 19th century.

[5] As discussed in Chapter 1, several concepts fundamental to statistics, including Legendre's method of least squares and Quetelet's curve-fitting, were published in non-English languages long before this period.

precisely how these textbooks were sourced, though the structure of Agresti's article seems to suggest that he found most of the works by looking through the "Reviews" section of contemporary journal publications.

On the other hand, Chapter 4 of Selden's *Inheriting Shame* details quantitatively and qualitatively the presence of eugenic ideology in American high school biology textbooks between 1914 and 1948 (roughly the span of the American eugenics movement). His corpus includes 41 textbooks, archived on "the shelves of the National Institute of Education Library's archive in Washington DC" and consulted physically during his study [85, p. 63]. Selden asked 8 questions of each book:

1. Is eugenics mentioned/cited?
2. Is eugenics recommended?
3. What words/phrases are used when referring to eugenics?
4. Are "inferior" families included as evidence?
5. Are "fitter" families included as evidence?
6. What hereditary traits are emphasized?
7. What prominent eugenicists are referenced?
8. What parts of the eugenic policy platform were referenced, if any?

Using updated statistical text mining technologies as described in the "Methods" section, I will replicate Selden's study on a set of higher-education statistics textbooks published between 1880 and 1970.

4.2 Data Collection and Methods

An initial list of statistics textbooks and pedagogical materials was gathered using the Library of Congress Catalog, searching for books under sub-classifications HA29-32 (Social Sciences–Statistics– Theory and method of social science statistics) and QA273-280 (Sciences–Mathematics–Probabilities. Mathematical statistics). Only textbooks written in English (including English translations of non-English works) were included. This list includes 719 distinct texts, which were cross-referenced with the HathiTrust digital archive. Of the 719 texts, 255 were available in full[6] through the HathiTrust Research Center datasets, and it is from these texts that the following analysis is conducted.[7] I abstained from including multiple editions of the same text, though text-

[6]Full text access for large-scale analysis is only granted for works in the public domain. Thus, the dataset is skewed towards earlier dates, as they are more likely to be in the public domain.

[7]There is a notable, but unavoidable, bias in using digitized resources: what gets digitized is largely due to modern, not historical, demand. Thus, books written by more famous authors like R.A. Fisher

books that were published in multiple volumes include all volumes, when possible.[8]

Initial data cleaning was done in Python 3.12 using the `ht-text-prep` package, published by the HathiTrust Research Center for processing text files downloaded via rsync from their databases [17]. Text mining was done in R 4.4.2 using the `tidytext` package [89] and the strategies described in the associated textbook [90]. Single-word tokens, bigrams, and trigrams were generated across the corpus, then matched to the phrases of interest to Selden. Each textbook in the statistics corpus was given a binary variable corresponding to each phrase, to match Selden's analysis. In other words, the proportions in the "Results" section refer to the percent of textbooks within the corpus that contain the indicated phrase. Due to the size of my corpus, it was not possible to manually check the context of each occurrence of a word or phrase. Thus, the frequency of a given word or phrase in the "Results" section does not indicate its frequency *in a eugenic context*, but rather its *overall* frequency.

4.3 Results

4.3.1 Comparison with Selden's Corpus

I will first discuss a direct comparison of Selden's results with my own. The following graphs contrast the proportional appearance of each phrase within the corpus, or, the number of textbooks in each corpus that contain the phrase divided by the total number of textbooks in the corpus. The numbers at the top of each bar give the number of texts in the specified corpus containing the phrase. In this section, I only compare my corpus and Selden's with regard to Questions 3-8 (as listed in Sec. 4.1), and provide a more robust view into Questions 1 and 2 in the following section.

The first two histograms below track the frequency of appearance of eugenic dog whistle[9] words and phrases. The dog whistle phrases that Selden traced through biol-

are much more likely to have their works digitized in the modern day, since modern historians are more interested in studying his publications than those of, say, Shepherd Dawson (1880-1935), a psychologist who worked in Pearson's lab during his graduate studies at King's College London. Dawson wrote *An Introduction to the Computation of Statistics* (1933), one of the textbooks that I was unable to obtain digitally. Dawson's book was not included in Agresti's article [1].

[8] The full list of textbooks included in the analysis is available at https://babel.hathitrust.org/cgi/mb?a=listis;c=1894270684 [44].

[9] A "dog whistle" is a political phrase that "sounds like ordinary speech, but… also conveys something problematic" [81, p. 330]. Common modern examples include " 'states' rights,' 'welfare,' 'inner city,' and 'law order' [sic]," which often are used to spread "a racist message without explicitly mentioning race." [81, p. 329, 330]. The phrase references a physical dog whistle, which produces a sound audible to dogs but not to humans. Galton introduced the dog whistle at the 1876 South Kensington Museum

ogy textbooks were by and large not found at all in the statistics textbooks (see Figure 4.1). However, searching for other phrases that are potential eugenic dog whistles did appear in higher frequencies, as shown in Figure 4.2. These phrases were sourced manually from Ruth Clifford Engs's *The Eugenics Movement: An Encyclopedia* [20]. Some of these words, like "disease" and "blood," are not necessarily unique to the eugenics movement, but were the subject of much eugenic concern. Others, like "breed," "fertility," and "pauperism," were more often than not linked to eugenic arguments.

The second set of histograms look at the ways in which eugenics was contextualized. The first, Figure 4.3, lists traits thought to be heritable, as discussed by Selden [85, p. 67]. Interestingly, "height" appeared with much greater frequency than in Selden's biology textbooks. This builds on Galton's tradition of tracking human height data. "Intelligence" and "industry" also appeared in a large proportion of the statistics textbooks. However, a drawback of the size of the corpus is the inability to manually check the context of these phrases. "Height" may well refer to the height of bars in a bar chart, or the height of non-human entities. "Intelligence" may or may not be spoken of in a eugenic context, and "industry," while an oft-spoken of quality in eugenic circles, more frequently refers to a field of business, like the lumber industry or automobile industry.

The second of this pair, Figure 4.4, examines the appearance of eugenic policy recommendations[10] in each corpus. Since the categories of policies listed in Selden's analysis are much broader than the words typically used to discuss eugenic policies, I have elected to split the categories into more specific phrases here. Thus, there is some incongruence with the graph in Selden's original work, but with the benefit of further specificity. Overall, Selden's textbooks showed a greater presence of eugenic policy recommendations than the statistics textbooks. The most common recommendations in Selden's study, differential birthrates and abortion, did not appear in the statistics textbooks. However, discussions of immigration and segregation were present in the statistics textbooks.

The third pair of histograms establishes the presence of "inferior" and "superior" families in each corpus. One of the hallmarks of the American eugenics movement was an obsession with certain families that had been singled-out as excessively "feebleminded" [85, p. 65]. None of these families — the Kallikaks, the Jukes, and the Ishmaelites — were mentioned in the statistics textbooks, as shown in Figure 4.5. Selden also presented certain lineages lauded in his textbooks for their seemingly hereditary genius, skill, and talent. Each of these families appeared at a lower rate in the statistics textbooks than in Selden's textbooks. However, as with several of the other his-

conference, detailing its efficacy through experiments he conducted at the London Zoo [11].

[10] Eugenic policy recommendations are discussed in further detail in Chapter 2.

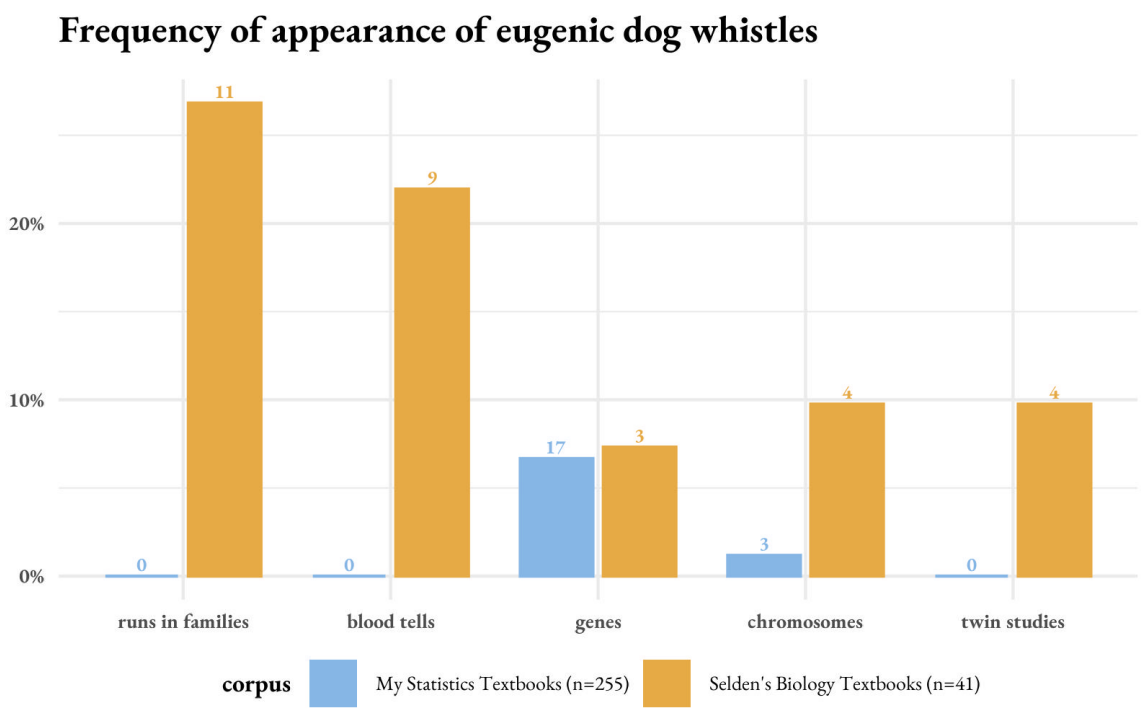

Figure 4.1: Percentage and raw count of textbooks in each corpus that contain the specified eugenic dog whistle phrases. Corresponds to Question 3 in Sec. 4.1 and Figure 4.2 in [85].

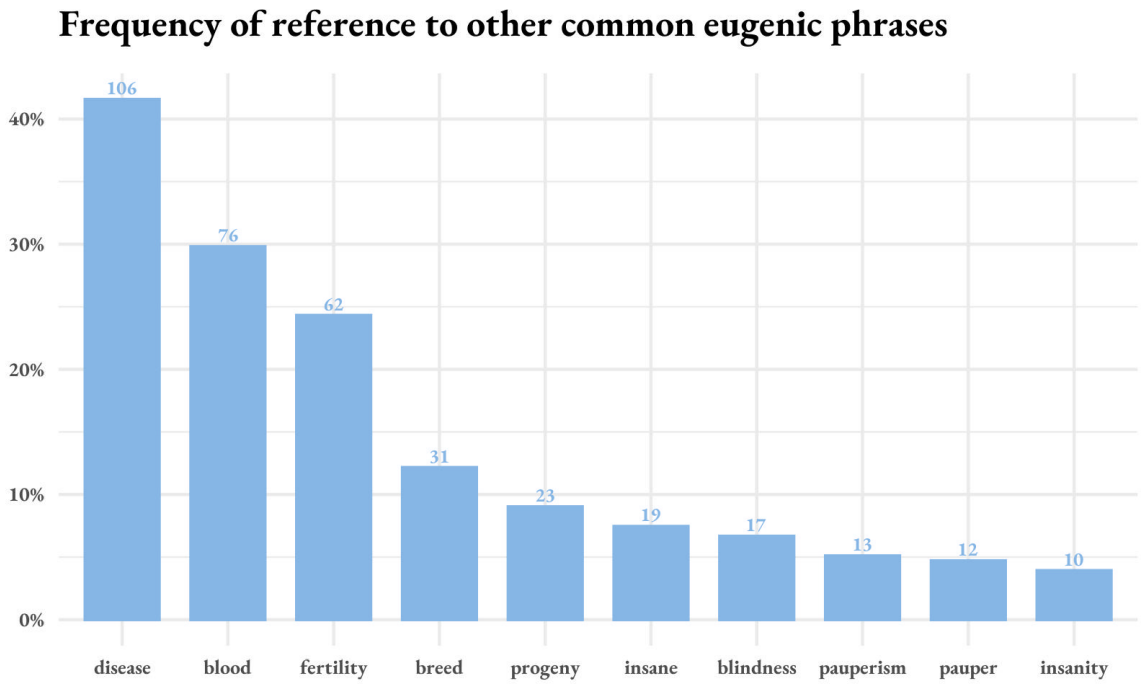

Figure 4.2: Percentage and raw count of supplementary eugenic dog whistles in the statistics textbook corpus.

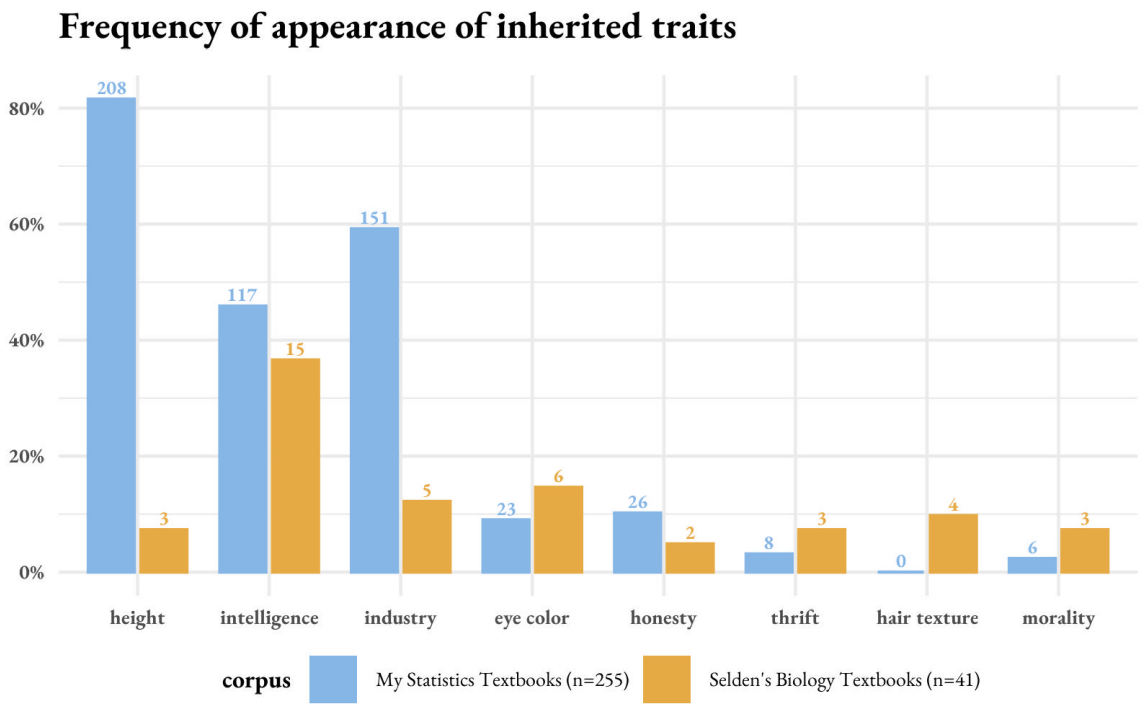

Figure 4.3: Percentage and raw count of textbooks in each corpus that contain the specified inherited traits. Corresponds to Question 6 in Sec. 4.1 and Figure 4.5 in [85].

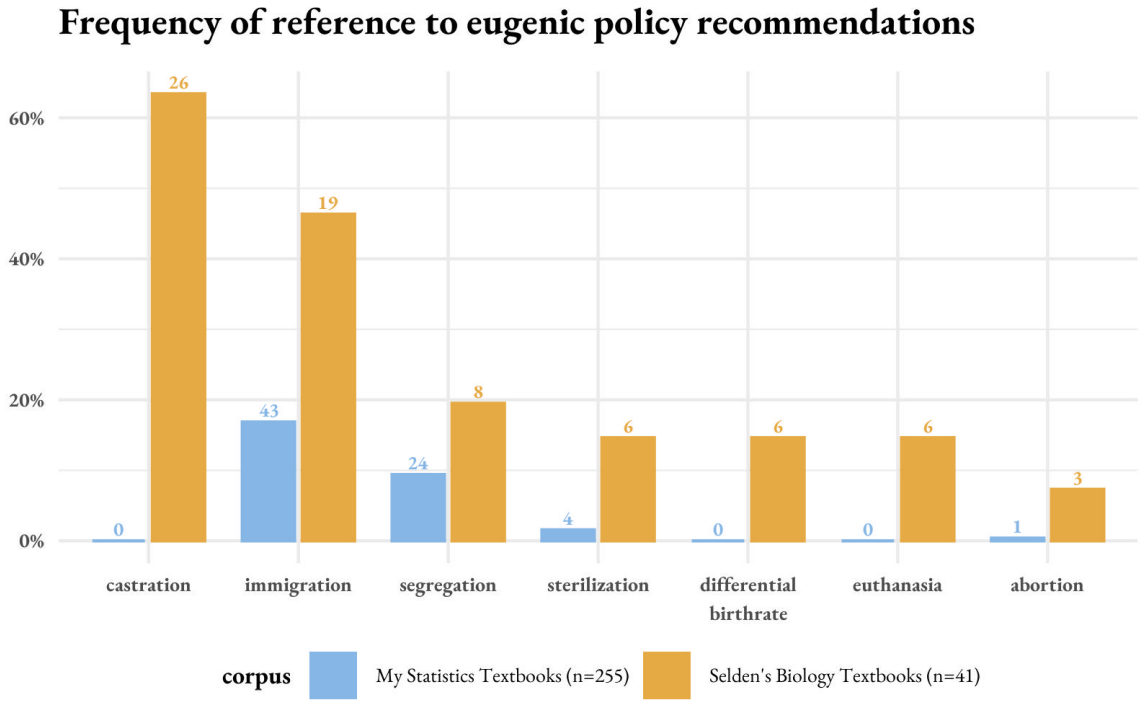

Figure 4.4: Percentage and raw count of textbooks in each corpus that contain the specified eugenic eugenic policy recommendations. Corresponds to Question 8 in Sec. 4.1 and Figure 4.7 in [85].

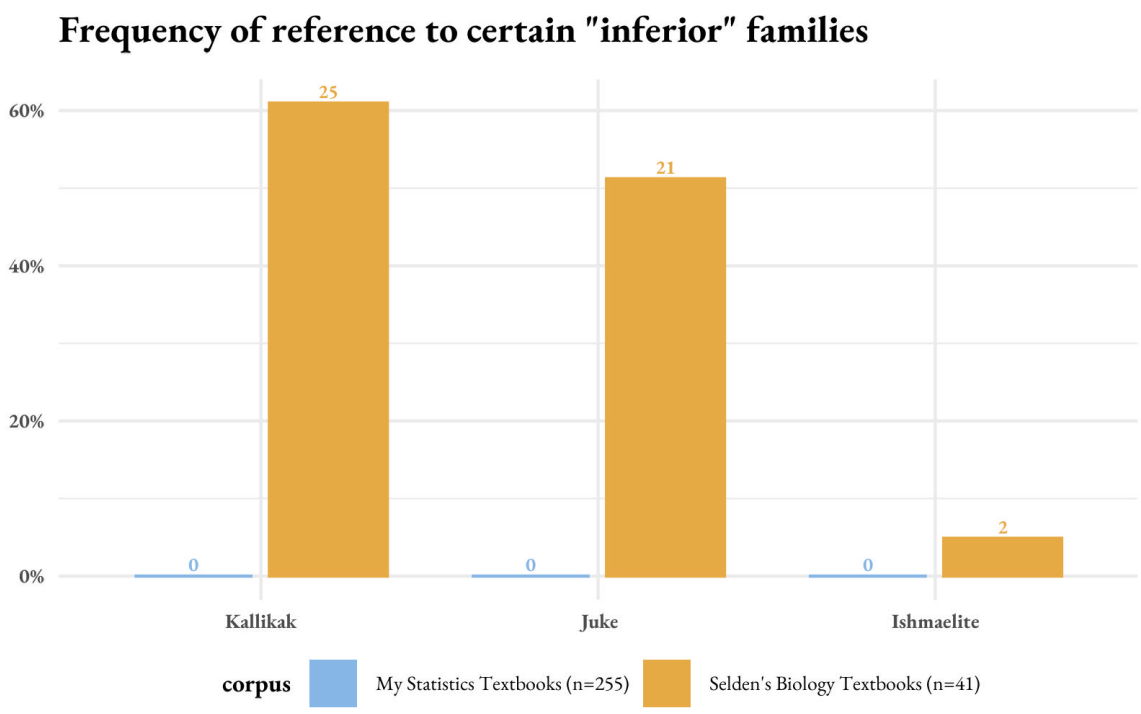

Figure 4.5: Percentage and raw count of textbooks that reference the specified "inferior" families. Corresponds to Question 4 in Sec. 4.1 and Figure 4.3 in [85].

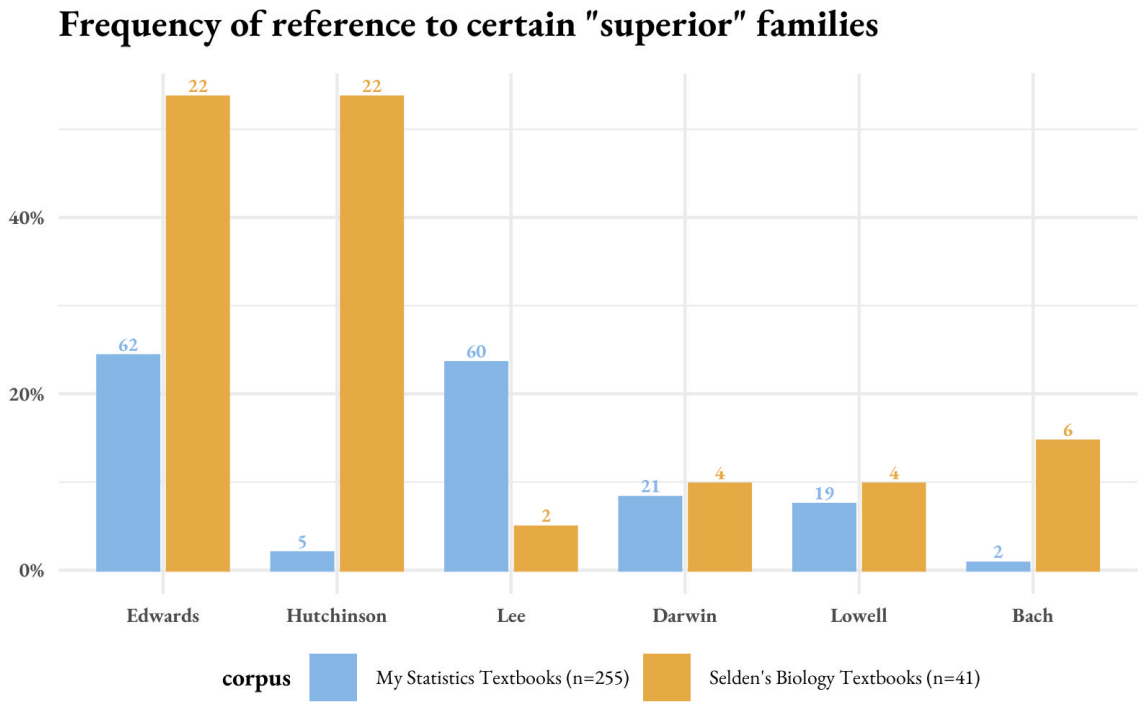

Figure 4.6: Percentage and raw count of textbooks that reference the specified "superior" families. Corresponds to Question 5 in Sec. 4.1 and Figure 4.4 in [85].

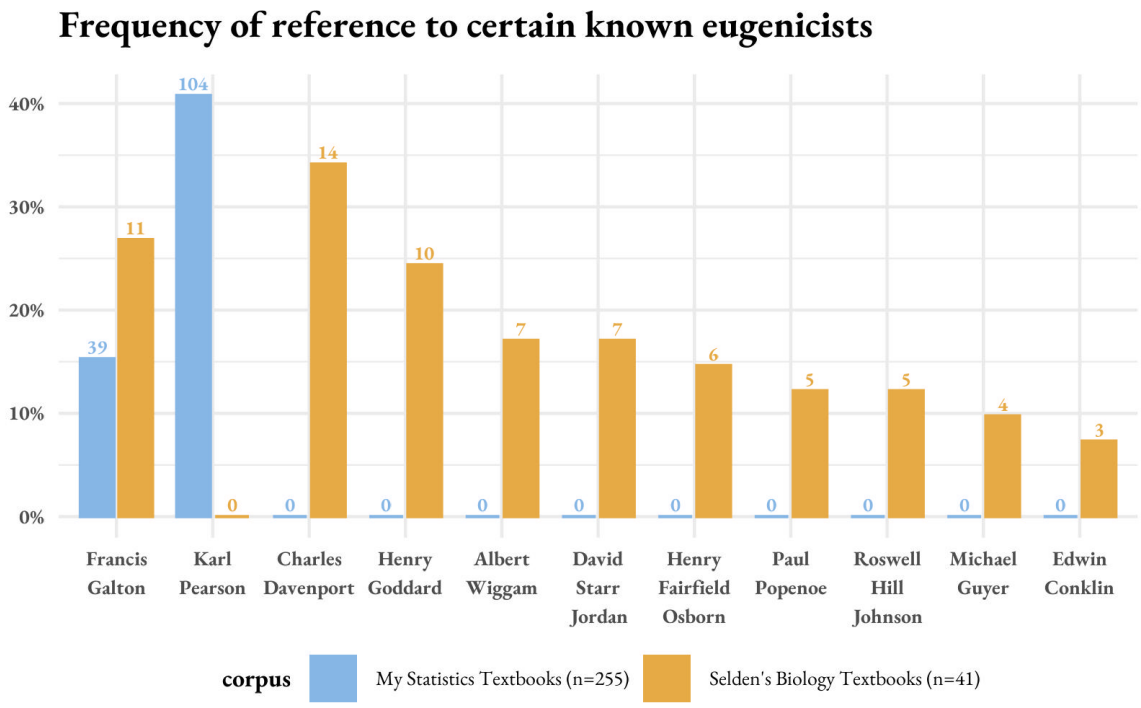

Figure 4.7: Percentage and raw count of textbooks that reference or cite the above leaders of eugenics movements. Corresponds to Question 7 in Sec. 4.1 and Figure 4.6 in [85].

tograms, it has not been confirmed that each occurrence of these names in the statistics textbooks refers to the a member of the family originally identified by Selden. For example, "Edwards" is a relatively common surname, and every person named Edwards does not necessarily belong to the same "superior" family. Regardless, taking the "inferior" and "superior" family histograms together, one can conclude that lineages and family trees were of more import in biology than in statistics. This corresponds to a division identified by M. Eileen Magnello between eugenics and biometry — the Galton Eugenics Laboratory concerned tracking and tabulating family trees and relations, while the Drapers' Biometric Laboratory did not work on such matters [58, 57].

The final histogram presents the frequency of reference to certain well-known eugenicists. Selden's original list was largely focused on strong voices in American eugenics movement, but included Galton due to his reknown as the founder of the movement. Since my analysis does not focus solely on American textbooks, I have also added

Karl Pearson to the list above, but otherwise it is unchanged. Pearson far surpasses all other eugenicists in terms of both percentage of textbooks that mention him and raw count. Meanwhile, Galton is the only figure to be mentioned across both corpuses. None of the pioneers of the American eugenics movement are referenced or cited in the statistics textbooks.

4.3.2 Frequency of Eugenics over Time

The first two questions asked by Selden concern the reference to and recommendation of eugenics as a concept. His study found that over 87% (36/41) of the biology textbooks included references to the eugenics movement, while over 70% (29/41) "recommended eugenics as a legitimate science" [85, p. 64]. However, a shortcoming of Selden's study was the agglomeration of 30 years of textbooks into one group. His study gives no indication of the distribution of frequency over time. If, for example, there was a fall in support for eugenics following World War II, one would have no way of knowing by reading his analysis. Thus, I include here frequency-over-time graphs to give some indication of possible time-based trends in the interaction between statistics and eugenics.

The first graph, Figure 4.8, displays the frequency of the word "eugenics" over time. The frequencies here are listed solely as raw counts. We see a peak in mentions of "eugenics" around the height of the organized eugenics movements in Great Britain and the United States. However, another peak occurs beginning in the late 1940s, around what Agresti calls the "[e]xplosion in mathematical statistics" [1, p. 674]. However, considering the size of the corpus ($n = 255$) and the total number of words it contains (2,3074,531), the total number of occurrences of the word "eugenics" is minuscule. Since this number was so small, I was able to manually find the context of each occurrence. The vast majority were found in the bibliographies of the textbooks, in the journal title *Annals of Eugenics*, established by Karl Pearson in 1925. These citations were usually to works by Pearson himself. Another common context was when stating the position of certain statisticians consulted in the writing of the textbook. For instance, William Palin Elderton thanks his sister Ethel Elderton, a member of the Galton Eugenics Laboratory under Pearson, for contributing to his book *Frequency Curves and Correlation* (1927). The one exception to the preceding contexts is Hugh H. Wolfenden's *Population Statistics and their Compilation* (1925), which begins by outlining its scope as "the statistical or collective study of human life" which contains as a subdivision "Human Eugenics, which investigates heredity from a scientific standpoint and is to a large extent the application of statistical method to genealogy" [100].

To contextualize the relative influence of eugenics, I also compare the frequency of

the word "eugenics" with more common words in the corpus: "table" and "data." Figure 4.9 shows the relative unimportance of the word "eugenics" in pedagogical statistics materials when compared with more general and directly applicable words.

4.4 Concluding Remarks and Future Directions

Overall, there was less of a eugenic presence in the statistics textbooks than in Selden's biology textbooks. At a rudimentary level, this shows that the discipline of statistics beyond Galton and Pearson were not motivated by eugenics in the same way. However, the precise text-matching used to analyze the statistics textbooks perhaps allowed instances of eugenic ideology to go undetected in the analysis. This method definitely resulted in the necessity for manual context-checking, which was not always feasible.

The modern methods used to extend Selden's analysis allowed for large-scale analysis of 255 different textbooks, and would be amenable to larger datasets if in the future more works entered the public domain. Furthermore, the dataset in its current state could be subject to more in-depth text-mining analyses, including those that use the term frequency-inverse document frequency (TF-IDF) statistic, n-gram correlations, or topic modelling algorithms.