TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

The Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Orthography and Biblical Criticism

A thesis submitted in partial fulfilment

of the requirements for the degree of Master of Science

by

Tomer Hasid

This work was carried out under the supervision of

Professor Amnon Ta-Shma

Submitted to the Senate of Tel Aviv University

March 2013

Acknowledgements

I wish to thank my advisor Professor Amnon Ta-Shma for all of his help, the creativity in solving every problem we encountered and for teaching me a lot about the process of research. I have studied a lot by working with him , and I'm sure it will give me a big advance for the future.

I also wish to thank Professor Nachum Dershowitz , for donating his part of the thesis, his ideas gave us a big push forward.

My great thanks also to Mr. Idan Dershowitz. His expertise in biblical criticism made this work possible and it was him who brought up the idea to use orthography to distinguish between the different sources.

Lastly , I would like to thank Professor Prof. Steve Fassberg for his great help in the paleological/neological classification.

Abstract

Biblical Hebrew exhibits considerable orthographic variability. A single word may be spelled in multiple ways—often within the same book. In this study, we set out to determine if these differences in spelling correspond in any way to scholarly theories regarding the authorship of the Pentateuch. Our results indicate that despite the tortuous editing processes and countless generations of hand-copied manuscripts, certain statistically significant correlations between orthography and the hypothesized sources remain.

This Work is a joint work with Proffesor Nachum Dershowitz and Mr. Idan Dershowitz.

Contents

Chapter 1

Introduction

The Pentateuch (Five Books of Moses) has been attributed to several major sources. In this study, we investigate whether there exists a correlation between these postulated Pentateuchal sources and variations in spelling in the received Masoretic biblical text.

We consider the source units about which there is broad–though hardly unanimous—agreement among Bible scholars, namely, the classic four-source division of the text into J, E, P, and D [Ł]. Additionally, considering the relative consensus regarding the existence of a Holiness Code with concomitant narrative, the source H was treated separately from P. In our analysis, we only consider words occurring in paragraphs for which there is some degree of consensus among scholars. We also compared genres, since different genres might have different conventions, and therefore separated the narrative and legal sections of these sources. [1] We also ignored the poetic sections, which are known to employ a quite distinctive language register.

Regarding orthography, we looked at the use of consonants to represent vowels, a practice that has changed over the centuries. Two types of writing systems existed in the ancient Semitic world: syllabic and alphabetic. The Canaanite languages—Hebrew among them—were generally recorded in alphabetic writing. [2] Paradoxically, this writing system took no account of vowels,[3] despite their substantial semantic and grammatical weight. With time, certain characters began to serve double duty, representing vowels, as well as consonants. These letters are known as *matres lectionis* ("mothers of reading"). The written representation of vowels increased from one century to the next, but it appears there was variation even within a single

[1] J's putative "Minor Book of the Covenant" was of insufficient scope for our purposes, so all of J is categorized as narrative.

[2] It appears the alphabet was in fact invented for use in a dialect of Canaanite. See [Ł].

[3] The term "abjad," rather than "alphabet," is sometimes preferred for such vowel-less systems.

time period.

While *matres lectionis* proliferated, a second process complicated matters somewhat. When a word's pronunciation evolved so that a particular consonant stopped being pronounced, the letter representing that now-absent consonant was not always written. For instance, יוצת (sans aleph) in the Masoretic text of Deuteronomy 28:57 is a defective variant of יוצאת (Amos 5:3, inter alia). It is, however, the latter form that reflects the earliest writing conventions, as it preserves the historical aleph that eventually ceased to be pronounced. For this reason, rather than the straightforward plene/defective dichotomy, we classify spellings as either "neological" (reflecting innovative orthography) or "paleological" (conforming to earlier norms).[4]

We have, then, two labelings to work with:

- By source and genre: J, E, E-law, P, P-law, D, D-law, H, or H-law. For simplicity, we will refer to these nine categories as "sources."
- By orthography: paleological or neological.

We apply a standard statistical test, Cochran-Mantel-Haenszel (CMH), to check whether there is a correlation between the two labelings. In other words, we seek to determine whether any particular source is more paleological or more neological than the others.

We began this investigation when one of us (I.D.), in his scholarly work, noticed what seemed like relatively consistent differences in spellings of several words between the Priestly material (P and H) and the other sources (e.g., אָרוֹן, *arōn*). But the results of the computer analysis were unforeseen. We have found, for instance, that Deuteronomic narrative and Priestly law exhibit wholly different spelling conventions, the former being considerably more neological. The results are summarized in Table 3.3 below.

We discuss possible approaches in the next section, which is followed by a description of our experiments and their results. Some statistical background is given in the Appendix.

[4] Since language does not evolve in a purely linear fashion, one branch might show certain signs of development long before a neighboring branch. Therefore, a "new" form can sometimes antedate "older" ones. Furthermore, register, personal style, and other factors can come into play. A late writer is often capable of writing in a pseudo-archaic style (sometimes referred to as "archaistic"), and an early scribe might opt for an informal spelling in certain circumstances. For more on this topic, see [Ł].

Chapter 2

Possible Approaches to the statistical problem

Assume we have two sources, $\mathfrak{A}$ and $\mathfrak{B}$, plus an orthographic classification, and would like to check whether the two classifications are correlated. There are several possible ways to approach this problem.

2.1 The Naïve Approach

A naïve approach is to count the total frequencies of neological and paleological syllables for each one of the sources and then run a χ^2 test for the resulting 2×2 table.

We believe the naïve test is not a good one. To illustrate, the word לא (*lō*) appears roughly 1205 times in the Pentateuch, only 12 of which are plene. [1] In an aggregated count, sources that often use the word will have a big push towards the paleological end of the spectrum, and those that use it less towards the neological one. This bias is so strong that it is likely to wipe out any real correlations for which we are looking.

More generally, if $\mathfrak{A}$ has a different word distribution than $\mathfrak{B}$, then it is possible that even when the two sources have identical spelling (and so the $\mathfrak{A}$/$\mathfrak{B}$ classification is independent of the neological/paleological classification) the naïve test would declare the two classifications as strongly correlated, simply because one source tends to use certain words that are spelled as neological (like עוֹלָה (*ōlā*)) more often. Working with aggregated data is therefore most likely to catch word distribution differences between sources, rather than spelling differences.

[1] We only consider occurrences in verses that are tagged as belonging to a single source.

2.2 Filtering

Andersen and Forbes [Ł] conducted an extensive automatized study of spelling in the Bible. Their approach was to classify all words in the Bible to 65 classes, based on grammatical form, vocalization and stress. Within each class (of the 65 they identified) they used the naïve approach described above, aggregating all words in a class and checking the χ^2 score of the 𝔄/𝔅 and plene/defective classifications.

As Andersen and Forbes use aggregated data (within each class), they still face the word distribution problem, and in particular their method is vulnerable to words like לא (*lō*) that appear frequently and mostly in one form. They tackled this problem with several ad-hoc filters and rules, such as filtering out words that almost always appear only in one form (see, e.g., [Ł, Chap. 10]). However, on a conceptual level, it seems that whatever set of filters is used, there is still the problem that differences in word distribution between different sources is interpreted as spelling differences.

In addition, it appears that Andersen and Forbes could not exhibit a conclusive relationship between stress and spelling (see [Ł, Epilog]), and this seems to undermine the rational behind dividing words into the 65 classes.

2.3 Our Approach

Our goal is to identify spelling differences even when each source may have a different distribution of words (e.g., legal texts would tend to use legal terminology). The standard statistical technique for doing this is the CMH test. The idea is to bypass the language distribution problem by having a 2×2 contingency table for each word in the language, describing the number of neological/paleological occurrences of the word in each source. The CMH test then combines the data from all the 2×2 tables in a way that gives weight to the statistical significance of the data in each table, but ignoring the frequency of the word in each source.

In fact, we choose to enumerate events at the finest possible granularity, that is we classify each syllable of each occurrence of a particular word in the text. For each syllable we have one stratum (in the statistical sense of stratified data) containing a 2×2 contingency table describing the number of neological/paleological occurrences of that syllable of the word in each source.

We think this observation, as simple as it is, is conceptually important and is crucial for getting sound statistical data on the problem. As a side effect of using the CMH test, we also

avoid ad-hoc filters and rules. Never the less, we ask the reader to bear in mind that it might be the case that there is some other hidden random variable that strongly affects spelling, which might better explain the results, and we hope our study will stimulate such study.

Chapter 3

Experimental Design and Results

3.1 Design

As explained before we have a stratum for each syllable in a word. We make use of the widely available tagging of the biblical text into word senses according to Strong number. For example, the Strong number of הָרִאשׁוֹן (*hārišŏn*) is 7223. For each word, we define its *base form* as follows:

1. Remove prefixes: הָרִאשׁוֹן (*hārišŏn*) → רִאשׁוֹן(*rišŏn*).

2. Reconstruct the word with all its syllables in a hypothetical maximally plene form: רִאשׁוֹן → רִיאשׁוֹן.

A stratum is then indexed by three components:

1. Strong number;

2. base form;

3. syllable number within the base form.

Thus, we have two strata of the form (7223,רִיאשׁוֹן,1) and (7223,רִיאשׁוֹן,2) for the first and second syllables in רִיאשׁוֹן.

We always consider two sources at a time (e.g., P-law and J), along with the classification of syllables by orthography. For each syllable, we count the number of times it appears as neological (paleological) in each source. For example, the counts for the *holam* syllable כוֹר

(*khor*) in the word בכור (*bekhor*) with Strong's number 1060 are:

Source	**P**	**P-law**	**D**	**D-law**	**H**	**H-law**	**E**	**E-law**	**J**
Neological	16	9	0	4	0	0	2	1	12
Paleological	6	0	0	3	0	0	1	0	3

Then, for each pair of sources we calculated 2×2 contingency tables, that is, for each stratum we keep only the columns belonging to the sources in question. We then compute the following statistics:

1. the χ^2 and p-value of the CMH test (see Appendix A.3);
2. the validity of the χ^2 test with the Rule of 5 (see Appendix A.5);
3. the common odds ratio (see Appendix A.7);
4. the $p = 1 - \alpha$ confidence intervals for the logarithm of the common odds ratio, taking $\alpha = 0.05$.

3.2 Results

Below are the p-values and the $\ln(\text{odds})$ values for the pairs of sources. The cells with tildes are those that failed to pass the Rule of 5.

	D	**D-law**	**E**	**E-law**	**P**	**P-law**	**H**	**H-law**
D-law	0.900							
E	0.073	**0.000**						
E-law	∼	0.198	∼					
P	0.323	0.848	0.777	∼				
P-law	**0.000**	0.087	0.588	∼	0.327			
H	∼	∼	∼	∼	0.445	∼		
H-law	0.296	0.804	0.482	∼	0.240	0.067	∼	
J	0.108	**0.033**	0.671	0.790	0.852	0.276	∼	0.184

Table 3.1: p-values for the pairs of sources.

In Table 3.2 below, the number in the cell (i, j) tells us how much is source i more likely to be paleological than source j. Roughly speaking, if the cell $(i, j) = 0.44$, then i uses the paleological form $2^{0.44} \geq 1$ more often then j. If it is zero they have the same frequencies, $2^0 = 1$. If it is negative, source i is less paleological, $2^{-0.44} \leq 1$

	D	**D-law**	**E**	**E-law**	**P**	**P-law**	**H**	**H-law**
D-law	-0.080							
E	0.460	1.192						
E-law	$\sim$	0.776	$\sim$					
P	0.263	0.118	-0.087	$\sim$				
P-law	0.818	0.503	0.210	$\sim$	0.181			
H	$\sim$	$\sim$	$\sim$	$\sim$	0.546	$\sim$		
H-law	0.516	0.161	-0.388	$\sim$	-0.389	-0.538	$\sim$	
J	0.351	0.521	0.107	-0.237	0.054	0.267	$\sim$	0.653

Table 3.2: ln(odds) values for the pairs of sources. If cell (i, j) is positive then source i is more paleological than source j.

Thus, D-law appears to be most neological and D is more neological than all sources other than D.

Lastly, we summarize, in Table 3.3, the p- and $\ln(\text{odds})$-values and also the χ^2 scores for those pairs of sources that—with a high level of confidence—display different orthographic styles:

Source Pair	χ^2	p-value	$\ln(odds)$
E vs. D-law	13.589	0.0002	1.1923
P-law vs. D	12.549	0.0003	0.8182
J vs. D-law	4.520	0.0334	0.5212
P-law vs. H-law	3.345	0.0673	0.5382
E vs. D	3.214	0.0730	0.4608
P-law vs. D-law	2.921	0.0873	0.5030

Table 3.3: Significant differences.

Notice that entries with very small p-values (like E vs. D-law and P-law vs. D) appear with odds-ratio above $e^{0.8} > 2.25$, showing that in these pairs of sources with high probability (as witnessed by the p-value) there is statistically strong correlation (as witnessed by expected odds-ratio). This kind of assertion can be made formal by calculating confidence intervals. Doing the calculation we see that with probability at least 0.95 the common odds ratio of E vs. D-law is in the range $[1.70, 6.35]$ and that of P-law vs. D in the range $[1.34, 3.81]$. The full table of confidence intervals (without H and E-law that are small and do not provide statistically significant data) is:

	D	**D-law**	**E**	**P**	**P-law**	**H-law**
D-law	$[0.51, 1.65]$					
E	$[0.97, 2.56]$	$[1.70, 6.35]$				
P	$[0.79, 2.11]$	$[0.58, 2.18]$	$[0.58, 1.44]$			
P-law	$[1.34, 3.81]$	$[0.91, 2.99]$	$[0.66, 2.28]$	$[0.84, 1.69]$		
H-law	$[0.73, 3.81]$	$[0.55, 2.49]$	$[0.28, 1.63]$	$[0.38, 1.20]$	$[0.32, 1.04]$	
J	$[0.94, 2.14]$	$[1.01, 2.79]$	$[0.74, 1.66]$	$[0.71, 1.55]$	$[0.82, 2.06]$	$[0.82, 4.50]$

Table 3.4: Confidence interval.

3.3 "Smikhut" Suffixes and orthography

One of the main grammatical characteristics of the Hebrew langauge is called "smikhut". Nouns have a construct state, "smikhut", which is used to denote the relationship of "belonging to". For example:

1. אָבוֹת - fathers;
2. אֲבוֹתֵינוּ - our fathers;
3. אֲבוֹתָם - their fathers;
4. אֲבוֹתֵיכֶם - the fathers of yours;

In all the above examples the different forms of the word fathers share the common holam syllable בוֹ (*bo*). Noticing this, we hypothesized that spelling would remain consistent enough to be statistically significant even when considering different forms of the same stem together. To our surprise, this does not seem to be the case. Bellow are the results (p-values) for the common syllables union:

	D	**D-law**	**E**	**E-law**	**P**	**P-law**	**H**	**H-law**
D-law	0.457							
E	0.325	**0.000**						
E-law	0.347	0.399	0.647					
P	0.774	0.810	0.538	**0.042**				
P-law	0.071	0.105	0.564	0.227	0.237			
H	$\sim$	$\sim$	$\sim$	$\sim$	0.416	$\sim$		
H-law	0.660	0.799	0.177	$\sim$	0.181	**0.049**	$\sim$	
J	0.778	0.056	0.966	0.599	0.812	0.579	0.460	0.119

Table 3.5: p-values for the pairs of sources after suffixes union. Notice that the result for D vs. P-law is not statistically significant any more.

To see why the results degrade, consider what happens when merging the different "smikhut" forms of the word יָדַיִ (*hands of*). The word יָדֶיכָה (*your hands*) may appear both plene and defective. The word יָדֶיהָ (*her hands*) always appears plene. Thus, in the test of

Section 3.2 the word יָדֶיהָ does not affect the results and can be ignored. On the other hand, when merging "smikhut" forms, יָדֶי appears both plene and defective and hence both יָדֶיכָה and יָדֶיהָ affect the results. In particular, the statistical significance of the plene/defective labeling is diluted.

On a more conceptual level, it seems that spelling is highly affected by the syllable placement within a word. For example, even for us as modern Hebrew speakers, it is clear that יָדֶיהָ cannot be written as יָדֶהָ, while both יָדֶיכָה and יָדֶכָה are valid. We believe this phenomenon is very common, and in particular explains the degrading of the results when "smikhut" forms are merged.

3.4 Can the picture be reversed? On using the statistical data to resolve paleological/neological labeling.

In all our tests, we designed the tests and obtained the paleological/neological labeled data, prior to running the tests themselves. Most of the human data was given with high confidence, for example, it is widely believed that גָּדַל is paleological and גָּדוֹל is neological. However, in certain cases, the experts were in doubt. For example, the labeling of אוֹן and עַמֶּיכָה was left undecided. In a few other cases, the experts decided on a classification after much hesitation. For example, שָׁלוֹשׁ was labeled paleological and שָׁלֹשׁ neological, even though the opposite labeling is not entirely ruled out. We stress again that to preserve statistical integrity we always firmly followed the experts' labeling.

Never the less, we feel that in some rare cases the statical data we collected seems to indicate the possibility that the experts' labeling might not be correct. The most prominent example for that is the holam syllable פוֹ (*po*) in the word צִיפּוֹר (*tzi-por*). The experts labeled צִיפּוֹר as paleological and צִפֹּר as neological, while our data seems to indicate the opposite (in D-law there are 2 occurrences of צִיפּוֹר and none of צִפֹּר, in P-law 2 occurrences of צִיפּוֹר and 9 of צִפֹּר, and according to our tests D-law is more neological than P-law). We stress again that to keep statistical integrity we always followed the experts' labeling.

Another "evidence" to the possibility that the experts' labeling might not be correct, is the following results for running the experiment with labelling of plene/defective only(instead of paleological/neological):

	D	**D-law**	**E**	**E-law**	**P**	**P-law**	**H**	**H-law**
D-law	0.423							
E	0.373	**0.006**						
E-law	$\sim$	0.318	$\sim$					
P	0.104	**0.029**	**0.042**	$\sim$				
P-law	0.000	0.000	0.161	$\sim$	0.657			
H	$\sim$	$\sim$	$\sim$	$\sim$	0.244	$\sim$		
H-law	0.377	0.131	0.435	$\sim$	0.457	0.061	$\sim$	
J	0.101	**0.053**	0.304	0.224	0.321	0.350	$\sim$	0.523

Table 3.6: p-values for the pairs of sources with labels of plene/defective only. Notice that the result for D-law vs. P-law is now statistically significant

Since for most of the syllables the plene form is the neological one we have reason to believe that these results show that for part of the syllables the final decision that the defective form is the neological one might be wrong.

Chapter 4

Discussion

Our results appear to be of potential interest to Bible scholars for several reasons.

- They suggest that the countless scribes who edited, expanded, and copied the text(s) that eventually crystallized into the Masoretic text did not change enough to obscure the characteristic spelling of individual units.
- Second, our findings open the door to new approaches in the critical analysis of biblical texts, as the value of orthography in such contexts has thus far been underestimated.
- Finally, the observation that Deuteronomic narrative is more neological in spelling than Priestly law may be of some value in the ongoing debate regarding the relative dating of P and D.

The simple statistical test we use cannot possibly disentangle the many authors of the Bible. However, it does produce some interesting results, that we hope would be combined with other data to shed light on the fascinating question of how the Bible, as we know it today, evolved.

Bibliography

[] F. I. Andersen and A. D. Forbes. *Spelling in the Hebrew Bible*. Biblical Institute Press, 1986. 2.2, 2.2, 2.2, A.4, A.4

[] A. Dmitrienko and W. Offen. *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Publishing, 2005.

[] S. R. Driver. *An Introduction to the Literature of the Old Testament*. Edinburgh, 1913. 1

[] J. L. Fleiss, B. Levin, and C. P. Myunghee. *Statistical Methods for Rates and Proportions, 3rd ed.* John Wiley, 2003.

[] R. D. Gill. The Cochran-Mantel-Haenszel test and the Lucia de Berk case. http:www.math.leidenuniv.nl/~gill/mantel-haenszel.pdf, 2007.

[] Orly Goldwasser. *Canaanites Reading Hieroglyphs, Part I – Horus is Hathor? Part II – The Invention of the Alphabet in Sinai*. Agypten und Levante, 2006. 2

[] H. Leckie-Tarry and D. Birch. *A Functional Linguistic Theory of Register*. Pinter, 1995. 4

[] N. Mantel and J. L. Fleiss. Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *American Journal of Epidemiology*, 112(1):129–134, 1980. A.4, A.5

[] J. H. McDonald. Cochran-Mantel-Haenszel test for repeated tests of independence. http://udel.edu/~mcdonald/statcmh.html.

[] Wikipedia page. Simpson's paradox. http://en.wikipedia.org/wiki/Simpson's_paradox. A.9

[] MathWorld-A Wolfram Web Resource. Hypergeometric distribution. http://mathworld.wolfram.com/HypergeometricDistribution.html.

[] J. Robins, N. Breslow, and S. Greenland. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, pages 311–323, 1986. A.7

[] W. Weinberg. *The History of Hebrew Plene Spelling*. Hebrew Union College Press, 1985.

Appendix A

A Statistical Discussion of the Problem

A.1 The Chi-Squared Test

The chi-squared test (χ^2) is a standard test seeking to *refute* the hypothesis that two classifications (e.g., neological/paleological spelling and sources $\mathfrak{A}$/$\mathfrak{B}$) of the sample space are independent. The assumption that the two classifications are independent is called *the Null hypothesis* and a high χ^2 score is evidence that the Null hypothesis is *false*.

	$\mathfrak{A}$	$\mathfrak{B}$	Total
Neological	200	50	250
Paleological	800	900	1700
Total	1000	950	1950

Table A.1: Observed frequencies.

	$\mathfrak{A}$	$\mathfrak{B}$	Total
Neological	128.2	121.7	250
Paleological	871.7	828.2	1700
Total	1000	950	1950

Table A.2: Expected frequencies.

Consider the 2×2 table in Table A.1. Given the total counts and assuming the Null hypothesis, the expected frequencies are given in Table A.2 (where, for example, the 128.2 in the neological-$\mathfrak{A}$ cell is calculated as $1000\frac{250}{1950}$, because we have 1000 samples from $\mathfrak{A}$, and each one of them should be neological with probability $\frac{250}{1950}$).

The χ^2 score measures the deviation of the observed values from the expected values (under the Null hypothesis) and is given by:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O and E are, respectively, the observed and expected frequencies tables. In the prior

example, the χ^2 score is 94.66. Notice that given the totals, any single element O_{ij} determines the rest. Thus, a 2×2 contingency table has only one degree of freedom—d.f. $= 1$.

The χ^2-score can be translated to a *p-value*, giving the probability of such a score or higher under the Null hypothesis and the d.f. For example $\chi^2 = 94.66$ translates to p-value $p \approx 2.26 \cdot 10^{-22}$ with the interpretation that the probability of getting such a score or higher under the Null hypothesis is smaller than 1 out of 10^{21} trials. Informally, it means that almost certainly there is some correlation in the sample space between being classified as neological/paleological and being classified as $\mathfrak{A}/\mathfrak{B}$.

A.2 Causality Inference

A high χ^2 score shows correlation between two properties A and B. It is tempting to interpret this as saying that property A is the cause to property B. For example, if high salary is linked with being female, one possible interpretation is that females are favored over males. This, however, is often not the correct interpretation, as we explain now. For example, in the above example it is possible that males and females are treated equally, but females prefer more demanding jobs.

Let us take, as a toy example, the sample space of all occurrences of the two words מֵאוֹת(*mēôṯ*) and עוֹלָה (ōlā) in the Pentateuch. (We are actually considering the second syllable of the word מֵאוֹת (*mēôṯ*), which is אוֹת (*ôṯ*) with *holam*, and the first syllable of the word עוֹלָה (*ōlā*), which is עו(*ō*) with *holam*.) The two classifications we choose are *narrative* (D, P, H, E, and J) versus *law* (D-law, P-law, H-law, and E-law) and neological versus paleological. (In neological spelling, the two words are spelled plene, whereas in the paleological case, the two are spelled defectively.) The result is Table A.3. We see that law is paleological, while narrative is neological, and the χ^2 score is about 117.97 with p-value about 10^{-27}, showing the results are statistically unquestionable.

	Law	narrative	Total
Neological	10	77	87
Paleological	93	10	103
Total	103	87	190

Table A.3: מֵאוֹת (*mēôṯ*) and עוֹלָה (*ōlā*): Observed frequencies

However, looking at the two words מֵאוֹת (*mēôṯ*) and עוֹלָה (*ōlā*) separately we get the two 2×2 tables in Tables A.4 and A.5. We see that the reason for the correlation between law/narrative and neological/paleological is mainly because in our simple example law tends to use more the word עוֹלָה(*ōlā*) which is most often spelled paleologically, and narrative tends to use more often the word מֵאוֹת *mēôṯ*) which is mostly spelled neologically. Thus, it is incorrect to say that law is more paleological, but rather there is another hidden variable—the distribution over the words in the source language—that explains the correlation.

	Law	narrative	Total
Neological	1	0	1
Paleological	93	9	102
Total	94	9	103

Table A.4: עוֹלָה (*ōlā*): Observed frequencies.

	Law	narrative	Total
Neological	9	77	86
Paleological	0	1	1
Total	9	78	87

Table A.5: מְאוֹת (*mēôṯ*): Observed frequencies.

If one knows which hidden variable most influences the results, then one can (and should) partition the dataset into different strata, analyzing each stratum on its own. The statistical test that analyses each stratum on its own and then combines the results is the Cochran-Mantel-Haenszel test, which we describe next.

A.3 The Cochran-Mantel-Haenszel χ^2-Test

The CMH test works with strata—which in our case are the 2×2 tables for each of the syllables—and its Null hypothesis is that the two classifications of the samples are independent for each (!) stratum. Suppose for the ith stratum we have the table:

	$\mathfrak{A}$	$\mathfrak{B}$	Total
Neological	X_i	$m_i - X_i$	m_i
Paleological	Y_i	$n_i - Y_i$	n_i
Total	$X_i + Y_i$	$(m_i + n_i) - (X_i + Y_i)$	$m_i + n_i$

Table A.6: Alternative notation for the frequency tables.

Under the null-hypothesis the two classifications are independent over the ith stratum and the random variable X_i should be hypergeometrically distributed[1] with parameters, as follows:

$m_i + n_i$: Population size
m_i : Number of neological words in the population
$X_i + Y_i$: Number of draws

[1] See Section A.10 for background on hypergeometrical distributions.

and therefore:

$$\begin{aligned} E(X_i) &= \frac{X_i + Y_i}{m_i + n_i} m_i \\ Var(X_i) &= \frac{(X_{i+}Y_i)((m_i + n_i) - (X_i + Y_i))m_i n_i}{(m_i + n_i)^2(m_i + n_i - 1)} \end{aligned}$$

The χ^2 statistic is given by:

$$\chi^2_{\text{MH}} = \frac{(|\sum_i X_i - E(X_i)| - \frac{1}{2})^2}{\sum_i Var(X_i)} \tag{A.1}$$

where the $-\frac{1}{2}$ in the numerator is a continuity correction, because the random variable X is integral.

Using the CMH χ^2 test for Table A.4 and Table A.5, we get χ^2 value about 0.554 and p-value about 0.456, which is very different than the aggregated score (which was 117 with p-value 10^{-27}) and shows that it is quite possible that the two sources have identical spelling habits.

A.4 A Remark on Filters

Previous work (like [Ł, Chapter 10]) applied ad-hoc filters to the data. For example, syllables that do not appear in both sources were filtered out as were syllables that are monochromatic (that is, have only one spelling), and, furthermore, these rules had to be extended for syllables that are "almost" entirely in one source or "almost" completely monochromatic (like the word לא(*lo*) that appears 1205 times defective and only 12 times plene in the Pentateuch). For a thorough discussion of these ad-hoc filters, see [Ł, Epilog].

The CMH test automatically filters out syllables that belong only to one source or monochromatic (because $X_i - E(X_i) = 0$ in these cases) and gives the correct weight to each one of the syllables when they are close to being monochromatic. As a result, there is no need to treat words like לא(*lo*) as a special case. Indeed another way of looking at the chi-squared statistic we have just defined is as the weighted average of differences between proportions, and the weight of the ith layer is larger for highly non-monochromatic syllables that appear a lot in both sources. For details see Section A.8.

As with all statistical tests one cannot deduce statistically significant conclusions from relatively small samples. To check whether the χ^2 value we calculated can be safely used for calculating p-values we use the *Rule of 5* thumb rule that was suggested by Mantel and Fleiss [Ł], as explained next.

A.5 Rule of Five

Consider Table A.6. Given the totals, and in particular, $m_i, n_i, X_i + Y_i$ we see that $X_i \geq 0$, $X_i \geq (X_i + Y_i) - n_i$, $X_i \leq m_i$ and $X_i \leq X_i + Y_i$.

Define

$$\begin{aligned} (X_i)_L &= \max(0, X_i + Y_i - n_i) \\ (X_i)_U &= \min(m_i, X_i + Y_i) \end{aligned}$$

and recall that

$$E(X_i) = \frac{X_i + Y_i}{m_i + n_i} m_i.$$

The rule of 5 requires that

$$\min\left[\left(\sum_i E(X_i) - \sum_i (X_i)_L\right), \left(\sum_i (X_i)_U - \sum_i E(X_i)\right)\right] \geq 5$$

See [Ł] for more details.

A.6 The Odds-Ratio of One Stratum

So far we discussed the *confidence* we attach to a suspected correlation. However, there is another important parameter, which is the *strength* of the suspected correlation. For example, assume a coin is "heads" with probability $0.5+1/10^4$. Thus, the coin is *slightly* biased towards "heads". If we throw the coin 10^7 times then we will notice this slight bias with huge confidence levels (that is, the probability our test shows a substantially different bias, and in particular the p-value, would be smaller than the number of particles in the universe). Thus, a small or negligible p-value does not necessarily mean a substantial correlation.

The odds-ratio is our guess at the strength of the correlation. Consider the 2×2 table with classifications neological/paleological and $\mathfrak{A}/\mathfrak{B}$ of Table A.7.

	$\mathfrak{A}$	$\mathfrak{B}$	Total
Neological	a	b	$a+b$
Paleological	c	d	$c+d$
Total	$a+c$	$b+d$	T

Table A.7: Observed frequencies for one syllable

Let A denote the event the sample is neological (and $\overline{A}$ paleological), and B the event the sample is from $\mathfrak{A}$ (and $\overline{B}$ from $\mathfrak{B}$). Define

$$\begin{aligned} \Omega_A &= \frac{P(B|A)}{P(\overline{B}|A)} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}, \text{ and} \\ \Omega_{\overline{A}} &= \frac{c}{d} \end{aligned}$$

Then the odds ratio is defined to be

$$\psi = \frac{\Omega_A}{\Omega_{\overline{A}}} = \frac{ad}{bc}.$$

Note that $\psi \geq 1$ implies $\Omega_A \geq \Omega_{\overline{A}}$. Thus, roughly speaking, $\psi \geq 1$ indicates positive correlation between A and B. Similarly, $\psi \leq 1$ indicates negative correlation and that A is less likely to occur when B occurs. The larger ψ is the larger the strength of the correlation is.

The confidence interval combines statistical knowledge about both the strength and the confidence of the correlation. A p-confidence interval tells us that with confidence p (that is, except for probability $1-p$) the odds-ratio is within some interval.

A.7 The Mantel-Haenszel Method for Multiple Strata

Mantel and Haenszel generalized the above to stratified classifications. They suggested the following estimator for the common odds ratio:

$$\widetilde{\psi}_{\text{MH}} = \frac{\sum_i \frac{a_i d_i}{T_i}}{\sum_i \frac{b_i c_i}{T_i}}$$

This turns out to be the weighted average of the odds ratio for each stratum. That is, if we let ψ_i be the odds ratio for the ith stratum, and the weights W_i be $W_i = \frac{b_i c_i}{T_i}$, then

$$\widetilde{\psi}_{\text{MH}} = \frac{\sum_i W_i \psi_i}{\sum_i W_i}.$$

The logarithm of the common odds ratio is a random variable, and furthermore, if the tests in each stratum are independent, it is normal. Assume we know its mean μ and its variance σ^2, then the probability we see a value larger than $\mu + z_{\alpha/2}\sigma$ is $\alpha/2$, where $z_{\alpha/2}$ is the value of cutting off the proportion $\alpha/2$ in the upper tail of the standard normal curve. The same reasoning holds for seeing a value $c \leq \mu - z_{\alpha/2}\sigma$.

The Mantel-Haenszel method estimates the common odds ratio and therefore also the logarithm of the common odds ratio. Robins et al. [Ł] give a formula approximating the variance of the logarithm of the common odds ratio. With it, one can estimate the p-confidence interval, that is, an interval $[L, H]$ such that for every value μ in the interval, the probability the correct mean is μ and yet we see our observed value is at least $1 - p$, and for every μ outside the interval, the probability is at most p. The confidence interval is

$$\left[\ln(\widetilde{\psi}_{\text{MH}}) - z_{\alpha/2}\sqrt{Var\ln(\widetilde{\psi}_{\text{MH}}))}\ ,\ \ln(\widetilde{\psi}_{\text{MH}}) + z_{\alpha/2}\sqrt{Var(\ln(\widetilde{\psi}_{\text{MH}}))}\right].$$

A.8 The CMH Test as Weighted Average of χ^2 Tests

Define

$$\bar{p}_i = \frac{X_i + Y_i}{m_i + n_i}$$

to be the proportion of $\mathfrak{A}$ in the i'th stratum, and let

$$\widehat{d}_i = \frac{X_i}{m_i} - \frac{Y_i}{n_i}, \text{ and,}$$
$$a_i = \frac{\widehat{d}_i}{\bar{p}_i(1 - \bar{p}_i)}.$$

The a_i capture the difference between proportions in the ith stratum. Define the "weight" of the ith stratum to be

$$w_i = \bar{p}_i(1 - \bar{p}_i)\frac{m_i n_i}{m_i + n_i}.$$

Then,

$$\chi^2_{\text{MH}} \approx \frac{\sum_i (w_i a_i)^2}{\sum_i w_i},$$

where the expression is only approximating the χ^2 score because we defined the χ^2 score (in Equation (A.1)) with a continuity correction.

A.9 Simpson's Paradox and the Delicate Pitfalls of Causal Inference

A natural assumption is that if the partitioned data show that, each stratum A is more neological than B, then the aggregated data must also show that A is more neological than B. This is, however, false.[2] The phenomenon is called Simpson's paradox and we illustrate it with an example taken from [Ł].

The University of California, Berkeley, was sued for bias against women who had applied for admission to graduate schools. The admission table for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it seemed unlikely to be due to chance.

	Applicants	**Admitted**
Men	8442	44%
Women	4321	35%

Table A.8: Berkeley's admission rate—all departments aggregated together

However, when examining the individual departments separately, it appeared , paradoxically, that no department was significantly biased against women. In fact, most departments had a small but statistically significant bias in favor of women. The data from the six largest departments are listed below.

Department	**Men**		**Women**	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Table A.9: Berkeley's applications for admission by department.

Looking at the data one can observe that women tended to apply to competitive departments with low rates of admission (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission (such as in engineering and chemistry).

The example illustrates that, when a hidden random variable is present (such as the different departments) and not taken into account, one can wrongly infer that there is causality between two classifications (like gender preference).

The conclusion is that even when we see significantly sound statistical correlations we should be very careful in our interpretation of the results. It is crucial to first understand the data, and correctly identify which variable most heavily influences the results.

[2] In fact, over $2 \times 2 \times 2$ with "random" data, it is false with probability about 1/60.

A.10 Hypergeometric Distributions

Suppose there are $N+M$ possibilities, N of which are "good" and M "bad". We take T samples without replacement and we let $x_i = 1$ if selection i is successful and 0 otherwise. Let X be the total number of successful selections, then

$$\begin{aligned} X &= \sum_{i=1}^{T} x_i. \\ P(X=k) &= \frac{\binom{N}{k}\binom{M}{T-k}}{\binom{M+N}{T}} \end{aligned}$$

and the mean and variance are:

$$\begin{aligned} E(X) &= T\frac{N}{N+M} \\ Var(X) &= \frac{TMN(N+M-T)}{(N+M)^2(N+M-1)} \end{aligned}$$