

Peer Review

Review of: "A Statistical Investigation of the Synoptic Problem"

Jacques Savoy¹

1. Computer Science Dept., University of Neuchâtel, Switzerland

This paper presents a study analyzing the authorship relationships between the first three gospels written by Matthew, Mark, and Luke. This is an interesting question. But even if this question has been studied under various perspectives, the current article proposes a new approach based on a linguistic background and a statistical procedure.

The abstract provides a good summary of the work. The question is well presented in the introduction. The main idea is to underline the wording relationships between these three gospels with passages that are very similar across the two/three gospels (a sequence of words). Such relationships between the synoptic gospels are known, but can we go further and can we detect the first author and the “copyists”? Maybe an example could show this question. In addition, the introduction signals the (established?) fact that Mark’s gospel was the first to be written. As explained by the author, the scholars did not suggest a common set of relationships between the two remaining gospels. For some, a source (the Q one) can explain part of those relationships. For others, instead of being limited to one additional source (Q), can we consider a second one?

This study is based on the HHB synoptic concordance. The author must confirm that the used material is Koine (or Ancient Greek) and not English. Moreover, it is important to underline that this language has a richer morphology than English (e.g., with grammatical cases denoted as suffixes). For example, one cannot have a single translation of “the” into Koine. This question can be relevant if the concordance is based on surface words (or word-types) and not lemmas (or entries in the dictionary). This aspect is not clearly mentioned in the paper. In addition, a small example of the concordance used in the study will help the reader (not necessarily a real one, and it could be in English).

The method is explained in plain English in the introduction. I don’t fully understand the two examples on Page 3. First, the number of words in common between X and not-X (or Y and not-Y) is rather small

(and ignoring the fact that “The” starts a sentence reduces the number to two). Second, the direction or the fact that X (or Y) is the first author is not clear. How can we deduce this? And at the end of Page 2, the idea is to assume that Mark is the first author.

2nd line on Page 4. “word occurrence”? Ambiguous term, word-type? Lemma?

Page 4: Good idea to ignore rare words (occurring once or twice). Because when analyzing word occurrence, we face LNRE (Large Number of Rare Events) distributions. Here we ignore them. Good point, and such evidence is too rare to support authorship attribution with some certainty.

The choice of the Poisson distribution is not well explained/supported. We can mention Mosteller & Wallace’s book, who also suggested a Poisson distribution for words (but they preferred the negative binomial (in part because the negative binomial distribution has a larger tail)).

My main concern is to assume that the variance is equal to the mean (in accordance with a Poisson distribution). I have selected the word ‘λέγει’ (he says) that occurs 2.2 times (per thousand) in Mark’s gospel with a variance of 5.4. (e.g., for other words (e.g., ‘τῶν’ in Matthew’s gospel), the agreement is good (mean: 10.1, var: 11.9)).

Page 5. Confirm that the synoptic categories are provided by HHB.

Page 6. Good point to ignore occurrences occurring in the test category when estimating the gamma values. However, it also means that we ignore a sample with at least one occurrence of the target word. And when the frequencies are relatively high, I don’t think it’s a real problem. However, when the occurrence frequency is lower, it could be a concern and renders the statistical procedure less authoritative.

Page 7. Using the MLE is a good idea (especially with not infrequent words). Specify also that beta must be between 0 and 1.

The likelihood ratio test makes sense. But this test is sensitive when comparing words having different means in the two samples. The resulting p-value could be small and significant (using the chi-square with 1 dof is correct). But if the two means are obtained with only a few occurrences, the reliability of the results is clearly lower.

Page 8. Top of the page. More precision is required. Previously, the relative frequency was computed according to the total number inside a category. Now, do we compute the mean over all categories?

Point 2. Why use the absolute frequency here? Is it to estimate the gamma_{predictor}?

To confirm. You just compute one beta value valid for all categories?

Taking account of the burstiness of word distributions is a good idea.

Page 11. Table 1. For each predictor category, how many words are present? Confirming an authorship relationship between two texts certainly requires more than a few words. In the current version, it is not clear if the p-values are obtained based on a single word (I don't think so), but it could be only a few.

I don't remember where, but the quasi-Poisson was not fully defined.

Starting with Section 5, the contents of the rest of this paper are related to biblical studies, which is not in my domain of expertise. Therefore, I'll not comment in detail.

Overall, the content makes sense and confirms that Mark's gospel must be viewed as the first one, and with at least one additional source (Q?), Matthew and Luke's gospels are related to Mark... Not fully sure that one can detect (with the proposed procedure) a strong and clear link between Matthew and Luke. But further studies must investigate this possible connection with other statistical or more traditional methods.

References

F. Mosteller and D.L. Wallace: Inference and Disputed Authorship, *The Federalist*. Addison-Wesley, Reading, 1964.

Declarations

Potential competing interests: No potential competing interests to declare.