On the Paper of Witztum, Rips and Rosenberg on Equidistant Letter Sequences in the Book of Genesis

Gil Kalai *

August 21, 1997

Abstract

In their paper [2] Witztum, Rips and Rosenberg (WRR) claimed to have statistically proved the existence of a hidden code in the Book of Genesis giving information about personalities living many centuries after the Book of Genesis was written. Their test was first conducted on 34 most prominent Jewish figures and another test was made on a second list of 32.

We give statistical evidence to the hypothesis that the significance in the second test of Witztum, Rips and Rosenberg was achieved via an optimization process in choosing the data, which was stopped when the significance level of the first test was met. It goes without saying that such a procedure is completely illegitimate.

Moreover, our results further suggest that the optimization process was done (at least in its final stages) by adding favorable appellations for the Rabbis until the addition of a single appellation moved the significance level beyond that of the first test. This compliments earlier findings of Bar-Natan, Bar Hillel and McKay which also indicate that optimization by choosing favorable appellations took place.

WRR informed in their 1987 preprint [3] a significance level of $1.29 \cdot 10^{-9}$ for the first experiment and $1.15 \cdot 10^{-9}$ for the second experiment. It turns out that these numbers are unreasonably close together. The gap between them is significantly small with respect to random divisions of the 66 Rabbis into two parts of 34 and 32 Rabbis respectively, with respect to random perturbations of the original lists, and with respect to the list obtained by moving to the second

Institute of Mathematics, Hebrew University, Jerusalem 91904 Israel.

list a certain Rabbi which was on the first list by mistake (according to WRR's criterion). The gap between these significance levels is considerably smaller than the typical affect of an addition of a *single* appellation to a single Rabbi.

The *square* of the ratio between the two significance levels reported by WRR is comparable to the typical effect of adding an appellation. This agrees with an optimization process of adding favorable appellations which stops when reaching the significance level of the first test.

Standing alone this finding gives strong statistical evidence that the work of WRR is deceptive and gives interesting insight into the nature of their deception. Our findings do not stand alone but complement the comprehensive study by Bar Hillel, Bar-Natan and McKay who studied various aspects of the two experiments conducted by WRR and reached the same conclusion.

1 Introduction

This paper contains some findings concerning the work [2] of Doron Witztum, Ilya Rips and Yoav Rosenberg (briefly WRR) in which they claimed to have statistically proven the existence of a hidden code in the Book of Genesis giving information about personalities living many centuries after the Book of Genesis was written. They claim to have shown significant proximity of Equidistant Letter Sequences (briefly ELS) of names and appellations of certain Rabbis w.r.t their (known) dates of death and birth. Their first test was conducted on a list of 34 most prominent Jewish figures and another test was made on a second list of 32. The first list of 34 Rabbis consists of all Rabbis whose entries in the *Encyclopedia of Great Men in Israel* by M. Margalioth is at least 3 columns long (such that either their date of birth or date of death is mentioned there). The second list consists of Rabbis having 1.5-3 columns in that encyclopedia.

WRR developed a statistical way to associate to a pair of terms a and b a certain "distance" $c(a,b)$ which is a rational number between 0 and 1 representing the quality of the most proximal ELS appearance of a and b. $c(a,b)$ can be regarded as a probability describing how lucky we are that a and b are found so close together in the Book of Genesis (see [2,3]).

WRR considered various appellations for each Rabbi and various standard forms for writing his known dates of birth and/or death. Then they computed all the distances between all pairs of

terms consisting of one appellation and one date for the same Rabbi and considered the significance of the combined list of numbers they got for all the Rabbis. (For more details, see [2], [3].) They found an extremely high significance level.

In their experiments WRR had to make a large number of choices. Some of these choices where fixed for the second experiment after being made in the first. Maya Bar Hillel discovered a large number of "degrees of freedom" concerning mainly the dates and various cases where the choices made by WRR were wrong according to their own criteria. Brendan McKay found that WRR did not implement the statistical test agreed upon with Diaconis but another one which inflated the significance level by a factor of 100. Dror Bar-Natan discovered vast degrees of freedom concerning the choice of the appellations. All these findings are described in a forthcoming paper by Bar Hillel, Bar-Natan and McKay [1]. Moreover, Bar Hillel, Bar-Natan and McKay showed that by choosing appellations for the 32 Rabbis of the second list in a certain way one can reach the same level of significance reported by WRR in the Hebrew translation of *War and Peace*. However, Witztum, Rips and Rosenberg claim that all their mistakes were innocent and have little effect on the final outcome, and that the choices they made were correct. They further claim their list of appellations was provided by an independent expert.

We give statistical evidence for the hypothesis that the significance in the second test of Witztum, Rips and Rosenberg is the result of an optimization process in choosing the data, which was stopped when the significance level of the first test was reached. Moreover, our results further suggest that the optimization process was carried out (at least in its final stages) by adding favorable appellations for the Rabbis until the addition of a single appellation moved the significance level beyond that of the first test. It goes without saying that such a procedure, which consists of manipulating the data to reach the desired goal, is completely illegitimate.

We will describe further facts about the WRR paper which are disturbing to its integrity and some directions for further research. In particular, the distributions of pair-distances (in both experiments) are friendly to the statistical tools used by WRR but do not support any reasonable interpretation of the original research hypothesis of a hidden text.

2 Some Examples

The case in hand is complicated and emotionally loaded. Moreover, the (hidden) assumption of divine intervention complicates things even further. Let us start, therefore, with simple examples which demonstrates the basic approach.

Suppose someone claims to be able to hit a globe hanging 200 meters away with a bow and arrows while blindfolded. You blindfold him, he shoots once, then after a while he shoots again, and then sends his son to bring the globe, and ...lo and behold! Both arrows are stuck in the globe on the equator, very close to each other. What the boy did not know, however, is that while his father was shooting, the globe was spinning round the axis through its poles very quickly, so that if an arrow hits the globe the longitude on which it lands is essentially random. Taking a closer look you calculate that the probability of two random points on the equator being as close as the two arrows is 1/100. Furthermore, the distance between the two arrows is about the closest the boy could have stuck the arrows without hurting his fingers if he was the one who stuck them in....

We observe here a phenomenon which supports a simple cheating strategy and is very unlikely to be found otherwise.

We will move now to an example which is quite close to the discussion in the next sections.

A researcher conducts a statistical test to check the hypothesis that there is a significant positive correlation between height and salary among people with academic education.

She interviews all people from one neighborhood in the city and sorts out the 90 with academic education among them.

Then she tests for each individual his height and his salary. She finds indeed a positive correlation and the significance level for her finding is 0.000132.

When she is asked by the referees of the paper to repeat her experiment she assigns one of her assistants to the job. He chooses another neighborhood, this time it is a neighborhood with many immigrants so finding those individuals with academic education is harder. He repeats the process (finds 80 persons this time) and proves again that there is positive correlation between height and salary with a significance level of 0.000120.

At this point the professor suspects that the assistant who wanted to ingratiate himself with her faked the results using the degrees of freedom he had. In particular, his precise criteria for a person to be qualified as having academic education could not be understood by her.

The similarity of the two significance levels raises the hypothesis that however the assistant tampered with the data his strategy was to do it gradually until the significance level of the second test passed that of the first test. The ratio between them is 1.1.

How can we test such a hypothesis? Assume that we have all the relevant data on the people considered as having academic education but not on those rejected.

Suppose we split the 170 chosen individuals in another way into two groups of people consisting of 90 and 80 people respectively. We can assume the significant positive correlation between height and salary will hold also for both these groups. However, there is no reason to assume that the proximity between the significance levels we observe in the two parts will typically be higher then that we observe in the original research. In fact, there are reasons to assume it will be typically smaller.

We can bound the significance of the proximity of the two significance levels by comparing it in a Monte Carlo experiment to the proximity observed by splitting the 170 chosen individuals to 90 and 80 at random.

Suppose we find out that with probability 1/100 the following event occur: *For a random splitting of these 170 people into two parts of 90 and 80, the ratio of the significance levels for the two parts is 1.1 or smaller.*

This would be an incriminating evidence since there is here a phenomenon that we could expect to occur if the assistant was tampering with the data in a certain way but was very unlikely to be found otherwise.

But we can make one further step. Suppose we suspect that the main degree of freedom of the assistant was in including or rejecting persons in his list and that he gradually added people with academic education to the second list which were favorable to the research hypothesis until he reached the significance level of the first test. In this case the proximity of the significance level of the two tests should be related to the effect of adding the *last person*.

If the typical effect of adding a single person to the list is compatible with the number 1.1, this will give an additional support to the cheating hypothesis. (As we will see later, being compatible means that typically adding a person with academic education to the list changes the significance levels by a factor in the neighborhood of 1.1^2.)

Several people were reminded by the Bible code case of the following story (possibly a tale), which I first heard from Maya Bar Hillel. The mathematician Poincaré bought loaves of bread from a certain bakery and after a while he complained to the police that the average weight he observed is 0.9 kilograms rather than the required 1 kilogram. The police intervened and since then all of Poincaré's loaves of bread were heavier than 1 kilogram. Six months later Poincaré was asked if the baker stopped cheating and his answer was that he didn't. He found statistical evidence to the fact that the baker kept cheating but that every day he was putting aside for Poincaré a loaf of bread which was heavier than 1 kilogram. Although all the loaves of bread Poincaré got were over 1 kilogram their distribution was significantly close to a normal distribution with average 0.9 kilogram cuts off at 1 kilogram.

Poincaré identified a simple cheating strategy of the baker so that the distribution of weights of his loaves of breads agrees with the distribution you expect if you assume cheating and is very unlikely to happened if there was no cheating.

All the examples of this section have the weakness that the statistical analysis was not made to give a priori predictions on new data but rather to study given data. With a good lawyer the baker may get off the hook. But the researcher, as soon as she got the picture, disqualified the experiment made by her assistant.

3 The significance levels in the research by Witztum, Rips and Rosenberg

Witztum, Rips and Rosenberg claim to have found an arrow sent by the author of the Book of Genesis, crossing thousands of years in its flight. But this arrow was not sent by the author of the Book of Genesis. The arrow was stuck by Witztum, Rips and Rosenberg themselves and they left

their fingerprints.

In their 1987 preprint [3] which presented the situation after the second test was carried out, Witztum, Rips and Rosenberg write in the introduction as follows:

"For the string G however, for the unperturbed sample we obtained $c(w, w')$ tending to zero with a probability against a null hypothesis of a uniform distribution that we estimate as $1.3 \cdot 10^{-9}$ for the first experiment and $1.2 \cdot 10^{-9}$ for the second (which gives the probability $1.8 \cdot 10^{-17}$ for the union of the samples)."

These numbers refer to the principal measure of significance used by WRR at that time for the two tests they made. The measure of significance used was later called the P2-statistics or the P2-score. In the Analysis Section of the same preprint the next digit is revealed so the numbers are $1.15 \cdot 10^{-9}$ and $1.29 \cdot 10^{-9}$, respectively.

The ratio between these two numbers is 1.1217. We will see that in view of the instability of the P2-statistics this ratio is extremely small. It is significantly small compared to random partitions of the 66 Rabbis into sets of 34 and 32 Rabbis respectively, it is significantly small compared to random perturbations of the original division of Rabbis, it is even substantially smaller than the typical effect of adding and deleting a *single* appellation to a single Rabbi.

The only explanation we can offer for this phenomenon is that there was an optimization process in the second experiment which stopped when the significance level of the first experiment was reached. Moreover, consider an optimization process in the second test which terminates with the addition of an appellation which brings the significance level beyond that of the first test. We can expect that the ratio of the significance levels of the two tests will be in the neighborhood of the square root of the effect of the last appellation. Indeed, the square root of the typical effect of adding an appellation is comparable to the ratio we witness.

It is worth noting that the astronomical significance levels cited above (from the 87 preprint [3]) are false due to wrong independence assumptions. A realistic way to measure the significance level suggested by Diaconis gives (for the second experiment) the value $1.6 \cdot 10^{-3}$. (See Section 7.)

4 Random partitions

I considered a random partition of the 66 Rabbis from the two tests together into one part of 34 Rabbis and another part of 32 Rabbis, and studied the distribution of the ratios of the P2 scores for the two groups and, in particular, how likely it is that such a ratio is smaller than 1.1217, the ratio of the numbers reported by WRR.

The distribution of the ratio of P2 statistics is quite interesting. The probability that this ratio is below 1.1217 is roughly 1/100. (More precisely it is 0.0092, but see the technical remark in Section 7.) Note that this is a direct Monte Carlo estimate and it does not rely on any probabilistic assumptions. The median value of the P2-ratio is roughly 700. The average is huge due to rare occurrences of very high ratios and I cannot estimate it. The average of the logarithm (with base 10) is roughly 3.3 .

As an illustration, if you move Rabbi number 7 (Rabbi David Ganz) from the first to the second list (and it is agreed that he was on the first list by mistake (see [1]) according the criteria of WRR,) the ratio in question changes to 4.1268.

Remark: We compute the probability that the ratio between the two P2-scores will be smaller than 1.1217. Of course, if we restrict our attention only to cases where the P2-score of the second test is smaller than that of the first test this further resuces the probability roughly by a factor of two.

5 The rationale behind the experiment and the basic consequences

We cannot expect the partition made by WRR to behave like a random partition. We should therefore explain what is the rationale for comparing its P2 score to a random partition. I will discuss separately two possibilities. The first is that the significant phenomenon described in the paper was the result of some optimization in choosing the data and the second is the original research hypothesis of WRR.

5.1 Assuming optimization took place

The explanation of WRR's results by optimization was suggested by various people and there is a comprehensive ongoing study by Bar Hillel, Bar-Natan and McKay concerning this possibility. It was previously believed that in both experiments there was some optimization (either intentionally or unintentionally) in order to improve the significance as much as possible. The above experiment suggests that, in fact, in the second test the results were improved until reaching the level of significance of the first test.

Indeed, without assuming this stopping procedure, there is no reason to believe that the P2 ratio of the original partition will not be smaller than that of a random partition and in fact, there are reasons to believe the contrary: in the two tests we have two distinct populations of Rabbis, the possible optimizations are different in the two tests (some parameters were determined by the first test) and the optimization skills are perhaps improved between them.

5.2 Assuming the original research hypothesis

Let's go now to the original hypothesis of WRR. Here too, not only is there no reason to believe that the two P2 scores will be more proximal to each other than those of a random partition but again there are reasons to believe the contrary.

Since one list is of the more important Rabbis and the other is of the less important ones, we can expect that the knowledge of the historical data will be different and also that the hidden biblical code will treat them in a different way. Thus, The two populations of Rabbis in the original partition are quite distinct, but in a random partition, where in each part we blend Rabbis from the two original lists, we can expect the two parts will be closer together.

5.3 Comparing with random perturbations of the original partition

The P2 ratio is not only significantly small with respect to random partitions but also with respect to random perturbations of the original partition. Thus, when you randomly replace k Rabbis on the first list with k Rabbis on the second list the probabilities of getting the P2-ratio of WRR or a smaller ratio are for $k = 1, 2, 3, 4, 5$ respectively: 0.035906, 0.022370 ,0.017372, 0.015322, 0.014150.

6 The effect of one appellation and a closer look at the stopping rule

One can say more on the possible stopping rule for the second test and where the P2 ratio we see came from. This is related to the effect of adding (or deleting) of *one* appellation. (Remember, each Rabbi has several appellations.)

The ratio of the P2 scores is still considerably lower than the usual effect on the P2 scores obtained by a single addition of a single appellation. But the *square* of the ratio obtained by adding one appellation seems quite close (still a bit on the lower side) to the ratio we experience.

Indeed in the second list there are 44 appellations whose deletion decreases the P2 score, 23 whose deletion increase the P2 score and 35 appellations which are dummies. (The "dummies" do not participate in any pairs of ELS and therefore they have no effect on the score.)

For 12 out of the 44 appellations whose deletion decreases the P2 score, the amount of decrease is smaller than the square of 1.1217 (roughly 27 %). This is the case for 9 out of the 23 appellations whose deletion increases the score (39 %).

This is compatible with an optimization process of repeated additions of appellations which stops when reaching a score which passes the P2 score of the first test. Indeed, in such an optimization process the ratio between the resulting P2 score and the P2 score of the first test is likely to be in the neighborhood of the square root of the effect of adding the *last* appellation. to the P2-score.

To see this pass to the logarithm of the P2-score and note that when we gradually add quantities and stop when we pass a threshold T then we can expect the difference between the outcome and T be in the neighborhood of the half the last quantity that was added.

(There are reasons to believe that smaller improvements will be delayed to later in the optimization, because of the dependencies of the pair distances it is more profitable to optimize in places where there are already excellent pair-distances.)

We cannot tell if this optimization was blunt cheating or if there was an innocent process (but totally wrong, of course,) were the criteria for including or rejecting appellations were formed using the distances observed in the Book of Genesis.

The realization of the large degree of freedoms concerning the appellations by Dror Bar-Natan (et al.) was a major turning point in understanding the work of Witztum, Rips and Rosenberg. Bar-Natan [1] discovered that the P2-score of the dates of the Rabbis in the second list with respect to the appellations of these Rabbis which were *not chosen* by WRR is significantly smaller than the the P2 score of the same dates with the same appellations randomly permuted.

In the example we gave in Section 2 Bar-Natan's finding is analogous to finding a significant *negative* correlation between height and salary for people who could have been considered as having academic education by the assistant but nevertheless were rejected.

7 Some explanations on the statistics and a technical remark on the distances

7.1 The P2 statistics

The P2-statistics describes the probability that the product of n *independent* random variables distributed uniformly in the interval $[0, 1]$ will be smaller than a number x.

The astronomical significance levels from the 87 preprints [3] are false (mainly) due to wrong independence assumptions. The pair distance $c(w, w')$ can be regarded as a function of the shortest ELS for w, the shortest ELS for w' and the distance between them. (The same remark applies to P1.) This creates massive dependencies among the pair distances. The continued use of false significance measures in the bible code activity is one of the reasons for the self-deception which is so characteristic of that activity.

The significance level reported in the paper [2] used a method suggested by Diaconis and is $1.6 \cdot 10^{-5}$. As pointed out by Brendan McKay [1] the authors did not implement Diaconis' suggestion but a rather different variant of it which inflates the significance level. The correct significance level for the second test if you implement the Diaconis test correctly is around $1.6 \cdot 10^{-3}$.

In the analysis section of [3] (and in [2]) WRR use also another statistics - later called the P1-statistics. The P1-statistics studies, using the normal approximation to the binomial distribution, the significance of the number of pair-distances whose value is smaller than 0.2.

7.2 The pair distances

A technical remark: The program computing the distances using the algorithm of WRR had repeated minor modifications (or debugging) and we do not have now the program used for the numbers appearing in the Statistical Science article and not in the earlier preprints.

At present there are two variants of the program. One variant (REAL_GEOMETRIC = OFF) is just a debugged (or slightly modified) version of the original one and the other variant (REAL_GEOMETRIC = ON) represents a slight methodological change which occurred after the paper appeared in Statistical science. The distributions of P2 ratios in both these variants are fairly close together (at least in the range of interest to us) and the 0.0092 figure is the probability of P2 ratio below 1.1217 in both of them. The individual P2 scores for the original partition did modify. The P2 scores in the modified version of the original concept is $1.741530 \cdot 10^{-09}$ for the first list $1.516102 \cdot 10^{-09}$ for the second and the ratio is 1.1487. The probability to be below this ratio is 0.0113. (For the new variant, which is irrelevant, the P2 score of the first list is $1.428 \cdot 10^{-09}$ for the second list it is $2.005 \cdot 10^{-09}$ the ratio is 1.40394 and the chance to be below it is 0.0278.)

We think 0.0092 is the right number to take and in any case the WRR preprint contains the original pair-scores so it is possible to insert them to the computer and use the original data.

8 The pair-distance distribution

8.1 the pair distances distribution for the two experiments separately

Consider now all the distances for all the pairs of appellations versus dates for all the Rabbis in each of the two tests made by WRR. There are 152 pairs for which the distances are defined in the first experiment and 163 in the second. And now consider the pair-distance histogram namely the histogram of the distances occurring at each experiment. (See, [2, p. 437] and [3, p. 4,5].)

One striking fact about the pair-distance distribution is that at least apparently they do not fit at all the suggestion made by WRR that there is a hidden text in the book of Genesis in which we can expect pairs of words which are "related" to be close together. The decreasing shape of the histogram even for "bad" distances does not seem to be supported by any reasonable hidden text

hypothesis. In particular what can be the reason for the rare appearances of very large distances (e.g. distances higher than 0.9)?

One could expect, for example, that the pair-distances histogram will be a combination of pair-distances for pairs which appear in the hidden text and pair distances for pairs which do not appear in the hidden text. Since three forms of writing the dates were chosen we can expect a substantial portion of pairs not to be in the hidden text. The histogram we see does not fit this possible description.

It seems that the pair distance histogram exhibits phenomena which are unfavorable to a theory of hidden text but are favorable to the main statistics chosen to verify the research hypothesis. This is not a good sign. It is like somebody tries to check the hypothesis that a certain university is lowering the academic standards for basketball players. He claims to prove this hypothesis by showing negative correlation between height and academic achievements. And then it turns out that this negative correlations is supported on small heights where there are no basketball players anyway.

In our view, the obvious explanation is that this is another sign of an optimization process which took place aimed at improving the P2 statistics.

It will be interesting to check a process that for a random ordering of all appellations of the two tests, pick them one by one adding to the list only those improving the P2 score (or perhaps choosing those with higher probability than the others) and stopping when the P2 score reaches (say) 10^{-9}. What will be the typical shape of the histogram of distances between pairs? This test can be conducted with respect to pair distances of the control tests given in [3], or for pair distance arising from distances in *War and Peace*. A simpler model would be to consider such a P2-optimization process when the data consists of independent random numbers with uniform distribution on $[0, 1]$ and at each stage 1-3 new such numbers are considered.

8.2 Similarity of the pair-distances histograms

The two histograms of pair distances from the two experiments look similar. (Indeed, the two graphs look somewhat linear.) If these two histograms can be distinguished in a statistically

significant way from histograms obtained from random partitions of the Rabbis, and if indeed they are significantly proximal then this will also be a very disturbing finding for the integrity of the WRR's paper. It seems that such a finding can be explained much better as a consequence of an optimization process than by some arguments related to the hypothesis of the research. This is a direction worthy of further study. The proximity of the two pair distances distributions looks also quite independent from the similarity in the P2 scores as the later is very sensitive to small perturbations of the partition.

The standard measure for the proximity of two distributions is the supremum norm of their difference, see [4, Ch. 14]. It would be interesting to estimate by a Monte Carlo experiment how significant is the proximity between the pair diatances distributions for the original partition of Rabbis with respect to random partitions.

9 A concluding remark

This work touches only on the tip of the iceberg in the amazing and saddening phenomenon of the "bible code" activity. The phenomenon is not so much about science or religion as it is about unlimited human ambition. I didn't touch at all the question of whether WRR proposed a genuine scientific phenomenon or theory and what should have been the right reaction of the scientific community to start with.

I always regarded Ilya Rips' claims that there is some hidden text in the bible as absurd yet harmless. The endorsement of Rips' work by quite a few mathematicians, the silence of the mathematical community at large, and Rips' and others use of these "codes" to gain historical and political insights changed matters.

My direct involvement was motivated by the recent publications which demonstrated that these codes can be quite harmful. In fact, it is quite possible that the "code" which was claimed to have "predicted" Rabin's assassination has been used for incitement or self-incitement against the late prime minister. The "objective", "scientific" recognition of this work makes these "bible codes" especially dangerous.

Acknowledgment

It is a pleasure to acknowledge the great help by Danny Braniss and by Leo Novik in programming. Brendan McKay supplied some raw data which was very helpful. I would like to thank also Maya Bar Hillel, Dror Bar-Natan, Ron Livné, Brendan McKay and Ilya Rips for helpful discussions on the statistical part of this work and many others who helped me greatly in presenting my ideas. Finally, I would like to mention the early work of Nahman Givoli who pioneered the critical study of the ELS activity.

References

[1] M. Bar Hillel, D. Bar-Natan and B. McKay, several works in progress.

[2] D. Witztum, E. Rips and Y. Rosenberg, On Equidistant Letter Sequences in the Book of Genesis, Statistical Science, 9 (1994), 429-438.

[3] D. Witztum, E. Rips and Y. Rosenberg, On Equidistant Letter Sequences in the Book of Genesis, preprint, 1987.

[4] S. Wilks, *Mathematical Statistics*, Wiley, New-York, 1962.