BS"D, 22 Cheshvan 5760 (Nov. 1st '99).

CONCERNING THE STATISTICAL TEST THAT WAS PUBLISHED IN OUR PAPER IN *STATISTICAL SCIENCE*

PART B

By Doron Witzum

Introduction:

In the article "CONCERNING THE STATISTICAL TEST THAT WAS PUBLISHED IN OUR PAPER IN *STATISTICAL SCIENCE"*, I pointed out the false assertion published by McKay *et al.* concerning the origins of this test. The false assertion is included in their paper "Solving the Bible Code Puzzle" [1], published in the May '99 issue of *Statistical Science.*

In response McKay *et al.* published an article on the internet [2] called "The origin of the permutation test", where they unsuccessfully attempted to conceal the fact that they had been caught red handed. But, as I will prove later, their arguments only strengthen our case. We will find McKay's *et al.* description of this issue very creative and imaginative but far away from reality. We will learn that such fairy tales are sometimes based on falsehoods and concealing of relevant data.

I. Which Test Was under Agreement?

1. According to the letters quoted by McKay *et al.* in their response on internet [2], it is perfectly clear that it is Diaconis who originally suggested using the permutation test. Concerning *how* to conduct this test McKay *et al.* point out two different methods: "Type A" as described in the correspondence of Prof. Aumann, and "type D" described in the correspondence of Prof. Diaconis.

In their article [1], McKay *et al.* claim that there was an agreement to use type D:

> "To correct the error in treating $P_{1\text{-}4}$ (that is, P_1, P_2, P_3 and P_4) as probabilities, Diaconis proposed a method that involved permuting the columns of a 32X32 matrix, whose (*i*,*j*)th entry was a single value representing some sort of aggregate distance between all the appellations of rabbi *i* and all the dates of rabbi *j*. This proposal was apparently first made in a letter of May 1990 to the Academy member handling the paper, Robert Aumann, though a related proposal had been made by Diaconis in 1988. The same design was again described by Diaconis in September (Diaconis, 1990), **and there appeared to be an agreement on the matter**." (Section 3, emphasis mine).

This is a falsehood.

McKay *et al.* themselves argue in [2] that the correspondence proves that there was disagreement and misunderstanding between the parties, and that Prof. Aumann stuck constantly to "type A" method, in complete contrast to what they wrote in [1].

2. McKay *et al.* continue in [1]:

> "However, unnoticed by Diaconis, WRR performed the different permutation test described in Section 2."

With this version of the story they create the impression that we deliberately deceived Diaconis and conducted an alternative experiment behind his back **contrary** to what had been agreed. This version of the story is distorted. Even the disjointed quotes brought in their own internet article [2] make it clear that this version is untrue.

Note: So far we have quoted McKay *et al.* who claim that had Diaconis suggested what they call "type D" method. But in our opinion, Diaconis who originated the idea of using the permutation test, only put forward the basic idea but never went into the details of our work. The impression that a different method was suggested is superficial, as Rips writes to McKay:

> "My guess is that Professor Diaconis probably did not look into the technical details of our work, and therefore he describes them in a general, vague and non-precise way." (From a letter of 4 April '97 which is quoted in [2]).

Despite this I do not wish to debate whether Diaconis had suggested a different method - "type D" – or not. We will assume for argument's sake that Diaconis did indeed suggest the alternative method D.

Diaconis's agreement to method A:

From the correspondence quoted by McKay *et al.* in [2] it is clear that even if Diaconis did suggest method D, Aumann clearly suggested method A, and in the final stage of the discussion Diaconis agreed to method A (the letter of agreement is found in Part A). Remember, in every negotiation it is the final agreement that counts.

We think that Diaconis's agreement was prompted by neither confusion nor inattention, but because he regarded it unimportant whether the experiment used method A or D. In any case it is clear that Diaconis furnished convincing reasons to create such an impression for seven years, so that no one including Aumann entertained any possibility that it was not so. Here are the reasons:

1. Diaconis, (together with all the referees of our article) received a draft of the paper which including a description of the experiment, before the experiment had been conducted. Assuming that he did not neglect his function as referee we may assume that he checked whether this was indeed the experiment he had agreed to.

2. In the next stage, after the experiment had been performed, he received the paper with the results. In the accompanying letter of 6 Dec. '91 (not quoted by McKay *et al.*) Aumann writes:

> "Enclosed is the paper of Witztum, Rips, and Rosenberg. The presentation was revised somewhat to make it clearer and take into account the comments of the people to whom I had previously shown it. **Needless to say, the test itself was not changed in any way; it is precisely the one to which we agreed in the summer of 1990" (**emphasis mine).

Had there been no agreement and had Diaconis thought that anything was remiss, he would have raised an objection. Especially considering that Diaconis was **against publication** of the experiment and had to invent a strange excuse to advise against its publication. According to McKay *et al.* Diaconis could easily have nullified the experiment by objecting that this was not the experiment he had agreed to.

3. In the summer of '92 Aumann and Diaconis discussed further projects related to our research. These discussions resulted in a letter of agreement on 28 August '92. In this letter it is clear that Diaconis does not dispute the fact that "the significance level is over 99.998%", and he makes no hint of any reservations about the method used - method A, not even concerning its use in future projects.

4. In November 97 McKay *et al.* wrote in their article in *Galileo*, no. 25 page 53:

> "Prof. Percy Diaconis, a world famous mathematician and statistician… suggested an alternative method to [WRR] which they used in their article published in ***Statistical Science.***" (emphasis mine).

In conclusion:

Even according to McKay's claims it is clear that:

a. Diaconis is the father of the idea of using the permutation test.
b. In the preliminary stage Diaconis suggested method D and Aumann recommended method A.
c. It was agreed to use method A (although McKay *et al.* claim that this was due to inattention on Diaconis's part).
d. Diaconis in his capacity as referee received a copy of the article both before and after the experiment was conducted and confirmed it.
e. During years of correspondence and discussions Diaconis never once objected that the experiment had utilized method A.

Finally after seven years, in '97, McKay comes along trying to persuade Diaconis that he didn't notice the difference between method A and D and that his whole agreement stemmed solely from inattention. This may be relevant to a psychological analysis of the thoughts and intentions of Diaconis, but it is totally irrelevant to the agreement between Aumann and Diaconis and the confirmation given by the latter as referee.

If McKay has objections – let him complain to Diaconis. But to create the impression that we acted unethically is **deliberate deceit of the public.**

II. Why Method A:

In chapter I, I clarified that the agreement was to use method A. I now wish to explain why this is the correct method to use, besides the fact that method D was never operatively defined. Diaconis' letter of 5 September '90 defines certain details of the experiment very vaguely and in particular fails to define the "distance" t (see Part A).

As McKay *et al.* wrote [1], the main purpose of Diaconis's suggestions were "to correct the error in treating $P_{1\text{-}4}$ (that is, P_1, P_2, P_3 and P_4) as probabilities".

Diaconis' objection was that to use these measures to direct evaluation of the significance is incorrect. Such use is based on the assumption of independence and uniformity, an assumption which Diaconis regards as wrong. The permutation test was designed to normalize the measures $P_{1\text{-}4}$, and therefore the way to do this is to compare the $P_{1\text{-}4}$ values of the original sample to the values of $P'_{1\text{-}4}$ of the "samples" created by permutations. This is exactly what method A does.

Method D, on the other hand, does not normalize the values of $P_{1\text{-}4}$ obtained from the original experiment. Instead it normalizes values of $P_{1\text{-}4}$ obtained for results of a totally different measurement (the "distance" t).

III. Distortion of method D and Concealing of Data by McKay *et al.*:

A. Distortion of method D:

In their article [2] McKay *et al.*write:

> "Aumann and Diaconis had agreed that a significance level of 1/1000 was a reasonable criterion for success. When WRR applied the permutation test they had designed themselves, they met that target easily. If they had performed the test that Diaconis wanted, on the same data, they would missed it. The failure of the experiment to pass the 1/1000 threshold would have greatly reduced its prospects of ever being published in a scientific journal."

We will see later that this claim is false.

McKay *et al.* deliberately wrote this passage very unclearly, and never even explained how method D is utilized. The answer to the question how is found in their article [1] where McKay *et al.* claim, in connection to Diaconis's method, that "using the average distance" is "the most obvious definition of his [Diaconis'] 32x32 matrix".

They claim all this explicitly in another article [3] (in connection to method B which will be explained in the next section):

> "This is a good place to note that (B) is the most natural interpretation of the experiment which WRR were asked to perform in 1990 by Persi Diaconis (on behalf of the journal to which their paper was first submitted). They failed to do so, but if they had the experiment would not have passed the 1/1000 milepost set for them. Whatever is the reason for it, the fact remains that [WRR] would quite likely not have been published if the Prof. Diaconis' instructions had been followed" (emphasis mine).

McKay *et al.* define method B in [4] as follows:

> "For each pair of persons p, p', compute one distance t(p,p') by averaging the defined values c(w,w') where w is in the first word-set of p and w' is in the second word-set of p'. If there are no such values defined, t(p,p') is undefined. For a permutation pi of the persons, define T(pi) to be the average over all p of the defined values t[p,pi(p)]. If there are no

> such defined values, T(pi) is undefined. The result will be the rank position of T(id) amongst all defined T(pi) for a large set of random permutations pi."

However, method B described here is completely incompatible to method D mentioned in Diaconis' letter of 5 September '90: Here the statistic T(pi) which is the average, **replaces** the two statistics P_1 and P_2 of method D! (P_3 and P_4 are repeats of P_1 and P_2 without the "rabbi" part of the sample, so altogether there are four statistics.)

Let us elaborate: Method B has three stages. For example concerning the second sample:

In the first stage we make an average *t*(p,p') of all the *c*(w,w') values of the pairs (name, date) relating to one individual and we get one number. If the set of *c*(w,w') values is empty, the average remains undefined. We repeat this for all the individuals in the sample and receive N numbers, according to the number of defined averages.
Step 2: We take the average of the N numbers and receive one number: T(id).
Step 3: We repeat the procedure on the samples obtained from the second sample through permutations, which pairs each individual with date connected to another individual. We then rank the T(id) amongst the T's received through the permutations. This rank is the significance level.

It is obvious that with this method it is impossible to use P_1 and P_2 (or P_3 and P_4). And this is a glaring contradiction to Diaconis' letter of 5 September '90, in which it is written that there is agreement to use these four statistics! (Note that this letter of Diaconis is in reply to Aumann's letter of 19 May '90, which defines these four statistics.)

Despite all this, McKay *et al.* maintain that Diaconis intent was to use the one and only statistic T which is totally different from the four statistics mentioned.

Therefore it is clear that the method forwarded by McKay *et al.* is their own invention, and was never suggested by Diaconis to Aumann. The absurdity is doubly apparent when one considers the comment of McKay *et al.* [2] on Diaconis' above letter: "Very little in his description is incorrect or vague." Who on earth can infer that a mutual negotiation and agreement about P_1, P_2, P_3, and P_4 is really about T which is totally incompatible with P_1, P_2, P_3, and P_4.

> How can McKay *et al.* have the audacity to assert that:
> "(B) is the most natural interpretation of the experiment
> which WRR were asked to perform in 1990 by Persi Diaconis".

The Definition of *t*.
Until now we discussed the definition of T in step 2 above. The definition of *t*, however, is relevant to step 1. Examining the correspondence between Diocinis and Aumann makes clear that there are no grounds to the claim that Diocinis

meant that *t* is the average. The only place Diaconis discusses a possible definition of *t*, (that is a numeric value representative of a set of *c*(w,w') values belonging to one personality) is in his letter of 3 August '88, and it is for the purposes of a preliminary investigation he conducted himself. And in this single place he writes "smallest" and not "average". Furthermore, in chapter IV we will explain why it is impossible to define *t* as an average if one wishes to use method D.

I would like to point out that I have only discussed all this to prove the emptiness of the claims of McKay *et al.*, and to draw attention to their deliberate and constant distortion of the facts. We ourselves, however, relied on what Prof. Aumann told us: That agreement was reached to use method A.

B. Concealing of Data:

Everything we said until now is bad enough, but it is not all.

After McKay *et al.* presented method B in the protocol of 17 April '97 (this protocal preceded [4]), my colleague Prof. Rips sent a letter to McKay on 1 May '97 explaining why he rejects method B, and in particular why the use of "average distance" is incorrect. He repeated this complaint in his critique of [3]. See [5] (He is quoted in chapter IV).

On 9 August '98 McKay published a document titled "Revisiting the Permutation Test". In this document he describes a new experiment, in which he uses a variant of method B (see definition of method B in previous section). In this variant step 1 defines *t* as: "the average of the logarithm of the defined c(w,w') values". Step 2 defines T as the sum of the logarithmic averages reached in stage 1. All this is supposed to answer Rips' above-mentioned objections. As McKay writes:

> "Use of the logarithm gives much stronger prominence to the word pairs with small distances' and in my opinion meets the objection of Rips that ordinary mean 'averages out' the alleged ELS phenomenon."

McKay presents his results:

> "For the WRR data, this method gives very respectable scores of 125/million for the first list and 8/million for the second. The combined list surely gives very much better than 1/million, but I have not computed it."

In other words McKay now reports a success of 8/1,000,000 for the second sample! He emailed this document on the above-mentioned date to his co-authors of [1]. However there is no mention of it in [1].

<u>Conclusions:</u>

A. McKay *et al.* know full well that the experiment's success was not due to deliberate choice of the randomization method, because McKay's method succeeded as well.
B. McKay *et al.* concealed the results of this experiment from the editors of ***St. Sc.***

This cover up is extremely significant: Once relevant statistical results are concealed there is little relevance to whatever else is presented.

We have no doubt that had the editors of ***St. Sc.*** known of this evasion they would not have published McKay's article.

Despite all this they have the audacity to write [1]:
"Nothing we have chosen to omit tells a story
contrary to the story here."

C. Would we indeed have failed to reach the threshold of 1/1,000?

We will now present the results of further experiments based on method D. If *t* is the numerical value that represents or sums up all the *c*(w,w') values of the pairs (name, date) relating to one individual, our original work had two such indexes to sum up the *c*(w,w') values: P_1 and P_2, and only them.

The use of $t = P_1$ is problematical when the size of the set of *c* values is very small, as would be expected with some of the individuals here, (there are even some sets with only 1-2 *c* values). But to give a complete picture we will use this possibility as well.

Accordingly, we must measure four statistics for the choice $t=P_1$ and four for $t=P_2$. Altogether 8 measurements. The best result obtained is for the statistic P_4 in the case $t=P_2$: **r**=0.000013.

The overall significance is therefore **r**=8×0.000013=0.0001 which is ten times better than the threshold of 1/1,000.

To complete the picture let us let conduct the experiment with the t=smallest which is the value chosen by Diaconis for his preliminary experiment. This is the sole example mentioned in his correspondence.

For t=smallest we receive: min $\mathbf{r}_{1\text{-}4}$=0.000153. which is a result 1.5 times better than the 1/1,000 threshold. (All the experiments were conducted with 10,000,000 permutations and with the same statistical seed used in the original experiment).

In conclusion: We have used for the definition of *t* all the ways to sum a set of *c(w,w')* values which were mentioned **before** the experiment, and we have found no foundation for the claim of McKay *et al.* that we would not have passed the 1/1000 threshold.

IV. Method D and the Average:

A. Concerning the usage of the average for summing up *c* values:

1. We are convinced that using an arithmetic average to sum up the results of experiments like ours is a fundamental error. On 1 May '97, about two weeks after Mckay *et al.* introduced the idea of method B Rips wrote to McKay:

> "Experiment B is absolutely unacceptable for me, and let me explain why. This research is oriented towards checking the claim that there is a hidden text in Genesis which is based on ELS's. We do not know what should be contained in this hidden text, so we make guesses. The input of each guess is a pair of words (w, w'). For each such pair of words we compute some functional *c*(w, w'). The functional *c*(w, w') was designed as to reflect

> some intuitive idea ("close meeting between ELS's", where "close" is understood for some cylindrical metrics on the text). To have a "small" value of *c*(w,w') means "success" (a close meeting between ELS's detected), otherwise "failure". Now we have to count the number of successes per number of guesses in order to decide whether we encounter a "remarkable" deviation from randomness. Both statistics P_1 and P_2 do it…
>
> Now what does the procedure of the experiment B? It AVERAGES the values of *c*(w,w'), in other words it punishes the successes for the failures. (For example, I would be very happy to have SYSTEMATICALLY a 1/100 per every 10 guesses; even such an impressive result would be AVERAGED OUT!)"

On 16 July 97 McKay *et al.* [3] replied and said among other things:

> "… while we acknowledge that (B) does not test for precisely the same phenomenon, it does test for something related…"

As we saw in the previous chapter, because of this discussion McKay found it necessary to make another experiment that

> "meets the objection of Rips that ordinary mean 'averages out' the alleged ELS phenomenon."

2. Let us give a further example why it is erroneous to use the average in our research. Let's say for simplicity's sake that every individual in the list has one appellation and three alternative date forms, in other words three (name, date) pairs. According to McKay *et al.* we need to take the average of three numbers (*c* values) for each individual. But it is possible to do the experiment differently: To conduct the experiment three times - once for each date form. McKay *et al.* did this and these are the results that appear in [1]:

Date Form	List 1	List 2
D M	0.165751	0.000017
BD M	0.000008	0.008844
D bM	0.008488	0.008804

The significance of this data is calculated as follows: We multiply the best value by three.
For the first list: r=0.000024.
For the second list: r=0.000051.

However, had we taken the average of the three numbers of each list the results would have been dramatically different:
For the first list: r=0.0581.
For the second list: r=0.0059.

Obviously, this is not the same as calculating an average for each individual by itself, but it suffices to demonstrate the absurdity of this approach.

B. Why is it impossible to define t=average for Method D?

In the previous chapter we showed how method B is completely incompatible with method D. In the previous section we basically explained why the use of the average is not appropriate to sum the results of our experiments. Now we will explain why it impossible to define t=average for the utilization of method D.

Remember: In method D we use the four statistics $P_{1\text{-}4}$ after the establishing of the "32×32 table of distances" through t. The two statistics P_1 and P_3 measure how many of the *c*(w,w') values are found in the segment (0,0.2], and give the probabilities for this.

We can now see that if we defined t=average, we cannot use P_1 and P_3 to utilize Method D. Let us give an example:

Let's say we have the following 10 values: 1/5, 1/5, 1/5, 3/5, 1/5, 4/5, 1/5, 5/5, 1/5, 2/5. The value of P_1 (or P3) is the probability that six out of the 10 values fall in the segment (0,0.2], and this is 0.00636. But the average of these ten numbers is 0.4 and is ***outside*** the segment (0,0.2].

An example from the second sample:
Individual no. 23 in the list has twelve values: 22/125, 10/115, 4/125, 1/125, 99/125, 5/115, 9/125, 56/125, 124/125, 78/115, 102/125, 19/125. The probability that seven values would fall in the segment (0,0.2] is 0.0039. But the average of the twelve numbers is 0.358, which is ***outside*** the segment (0,0.2].

In other words: The premise of our research expects an accumulation of *c*(w,w') values in the segment (0,0.2], and it is exactly the task of P_1 to detect it. But we can expect ***in advance*** that the passage from individual *c* values to arithmetic averages, will skew the results outside the segment (0,0.2]. A correct calculation requires finding the right transform of the bound 0.2 for the passage from *c* values to arithmetic averages. But in our case the calculation is even more complex because moving the 0.2 bound to the right also improves the results in the case of the individual *c* values.

As far as we are concerned we are convinced that Diaconis was not interested at all in the details of the experiment but only in its main principle.
However, anyone who claims that Diaconis suggested taking the t=average in the framework of the D method, is accusing him of purposely inventing a procedure knowing in advance that it would ruin the experiment. From the data we sent him it could be inferred that checking the accumulation of the arithmetic averages in the segment (0,0.2] would lead to the loss of any significance. We do not suspect Diaconis of proposing such an "experiment" that from the data he already had in hand, he could know that it would surely fail.

The data that was sent to Diaconis before he suggested the permutations test was the *c*(w,w') values for the second list. From this data alone one can calculate the average for each individual, and also know how many of these averages will fall in the segment (0,0.2] and what the probability for this is. For P_1 the probability is 0.92 and for P_3 the probability for this is 0.82. Thus from this data one could determine that the experiment would fail under such conditions even before the permutations test was conducted.

V. Concerning the Claim that Our Success Resulted From Use of Method A:

In an article [6] published in a scientific journal and in an accompanying article on internet [7], I demonstrated how a randomization method different from that of Diaconis is applied to the second sample. This alluded to our RPWL method (=Randomization by Permutations of Words' Letters) which was first applied in [8] for samples of "heading" type, samples that could not be treated using the permutation test of the type suggested by Diaconis. All McKay's objections concerning method A are irrelevant for RPWL, and the significance obtained was far better: r=0.00000188.

Until this day McKay *et al.* have not related to these results, nor have they shown any fault in this method of measurement. In their article [1] in ***St. Sc.,*** which is a review article, they criticize the original measuring method and bring their own replications, but they ignore this important replication and also the replication of McKay himself (see above ch. III section B). Both replications are relevant to the question whether the original results we received resulted solely from the measuring methods we used.

And after all this they have the audacity to write there:
"Nothing we have chosen to omit tells a story
contrary to the story here."

Bibliography

1. B. McKay , D. Bar-Natan, M. Bar-Hillel & G. Kalai, *Solving the Bible Code Puzzle*, Statistical Science, Vol. 14, No. 2, 150-173.
2. B. McKay & G. Kalai, *The Origin of the Permutation Test*, on web site: http://cs.anu.edu.au/~bdm/dilugim.
3. D. Bar-Natan, A. Gindis, A, Levitan, B. McKay, *The New ELS Tests – A rejoinder*, July 16, 1997, at the above web site.
4. D. Bar-Natan, A. Gindis, A, Levitan, B. McKay, *Report on New ELS Tests of Torah*, 29 May 1997, at the same web site.
5. E. Rips, *Preliminary Analysis and Comments on the Report of New ELS Tests*, June 19, 1997.
6. D. Witztum, *Concerning the "REMEZ" in Equidistant Letter Sequences (ELS's),* BDD, Journal of Torah and Scholarship, Bar-Ilan University Press, No. 7, Summer 1998 [in Hebrew]. Available at: http://www.torahcode.co.il/pdf_files/pub/bdd.pdf.
7. D. Witztum and Y. Beremez, *The "Famous Rabbis" Sample: A New Measurement*, available at our web site: http://www.torahcode.co.il (first version May 1998, updated Sept. 1998). Available at: http://www.torahcode.co.il/english/pdf_files/new2e.pdf.
8. D. Witztum, E. Rips, Y. Rosenberg, "*A Hidden Code in Equidistant Letters Sequences in the Book of Genesis: The Statistical Significance of the Phenomenon,"* (in Hebrew, preprint, Spring 1996).