EISENBRAUNS
Offprint from:
Rethinking Israel: Studies in the History and Archaeology of Ancient Israel in Honor of of Israel Finkelstein

Statistical Inference in Archaeology: Are We Confident?

Arie Shaus, Barak Sober, Shira Faigenbaum-Golovin, Anat Mendel-Geberovich, David Levin, Eli Piasetzky, and Eli Turkel
Tel Aviv University

Introduction

We deal with the general issue of handling statistical data in archaeology for the purpose of deducing sound, justified conclusions. The employment of various quantitative and statistical methods in archaeological practice has existed from its beginning as a systematic discipline in the 19th century (Drower 1995). Since this early period, the focus of archaeological research has developed and shifted several times. The last phase in this process, especially common in recent decades, is the proliferation of collaboration with various branches of the exact and natural sciences. Many new avenues of inquiry have been inaugurated, and a wealth of information has become available to archaeologists. In our view, the plethora of newly obtained data requires a careful reexamination of existing statistical approaches and a restatement of the desired focus of some archaeological investigations. We are delighted to dedicate this article to Israel Finkelstein, our teacher, adviser, colleague, and friend, who is one of the father figures of this ongoing scientific revolution in archaeology (e.g., Finkelstein and Piasetzky 2010, Finkelstein et al. 2012, 2015), and wish him many more fruitful years of research.

The Legacy of New Archaeology

The last major methodological attempt to formulate a positivistic statistical approach to archaeological research can be attributed to the "Processual," or "New," school of archaeology, which has profoundly influenced the form and content of archaeological studies since the 1960s. The proponents of the New Archaeology suggested that archaeological data can be analyzed on a deeper level, beyond the mere description of unearthed artifacts. Most of the advocates of the new theory were oriented toward the social sciences—in particular, anthropology. Indeed, this fact is epitomized by the title of one of its founding articles, "Archaeology as Anthropology" (Binford 1962). The New Archaeologists imported sociological and anthropological tools and formulations in order to arrive at conclusions concerning ancient societies and, hopefully, deduce laws affecting humanity in general.

The debate regarding the scientific (or "scientific") nature of processual archaeology, which began shortly after its appearance (Flannery 1982), is still ongoing. This also concerns its proposed improvements and substitutes, often collectively labeled

"post-processual archaeology" (e.g., VanPool and VanPool 1999; Hutson 2001; Arnold III and Wilkens 2001; VanPool and VanPool 2001). The existence of such a lively methodological dispute may be explained by the fact that the premises and the means of the New Archaeology pertain to the "soft," rather than the "hard," sciences and therefore can often be questioned.

Recent years have seen a rapid growth in employment of methods pertaining to the exact, earth, and life sciences in archaeological practice (e.g., Finkelstein et al. 2015). In our view, the application of such hard sciences in archaeology necessitates a revision in its underlying theory and the types of questions it attempts to address. In the current paper, we will concentrate on statistical methods that can be used by archaeologists. In particular, we shall contrast descriptive statistics, commonly applied in archaeology, with statistical inference methods (see details and examples below). It is our conviction that certain techniques of statistical inference should become the preferred *modus operandi* of archaeological inquiry, guiding its research questions and methods. In fact, *the possibility of utilizing statistical inference methods may present the core requirement of modern, truly scientific archaeology.*

The last statement echoes the opinion of the late Sir Karl Popper (1960), who anticipated difficulties in the application of "quantitative methods, and especially methods of measurement" in the social sciences. Nevertheless, he admitted that

> Some of these difficulties can be, and have been, overcome by the application of statistical methods, for example in demand analysis. And they *have to be overcome* if, for example, some of the equations of mathematical economics are to provide a basis even of merely qualitative applications; for without such measurement we should often not know whether or not some counteracting influences exceeded an effect calculated in merely qualitative terms.

Popper warns that

> merely qualitative considerations may well be deceptive at times; just as deceptive, to quote Professor Frisch (Frisch 1933), 'as to say that when a man tries to row a boat forward, the boat will be driven backward because of the pressure exerted by his feet'.

In other words, only accurate quantitative data, supplemented with sound statistical analysis, may counter the peril of unsubstantiated qualitative conclusions.

Descriptive Statistics versus Statistical Inference

In general, statistical methods can be roughly divided into descriptive and inferential (or inductive). *Descriptive statistics* deals with summarizing the assembled information. For instance, given a sample of ceramic sherds found in a survey, we can count sherds and derive relative proportions of each type—for example, 75% of Chalcolithic and 25% of Late Byzantine pottery. One may be tempted to derive conclusions based on these summaries—for example, that the site was larger during the Chalcolithic period than in the Late Byzantine period. This can be seen as a deduction of facts based on statistics, but this judgment is interpretive rather than statistical. In fact, this kind of reasoning can often be critiqued, since other factors may be taken into consideration, possibly hindering the conclusion (e.g., the total number of sherds, the relative length of the periods, occupational duration of each

site within the period, the quality and the uniqueness of the ceramics, preservational factors, as well as survey parameters).

On the other hand, *inferential statistics* supplements the raw summary with a score, allowing one to quantify the certainty of the results or at least to choose the most likely "scenario," after articulating some relevant assumptions. We provide a short summary of three common procedures of statistical inference:

- *Confidence interval* estimation: In this procedure, several empirical results are taken into account in order to estimate a required statistical parameter. Additional information, either pre-existing or estimated from the sample itself, is then used in order to establish a *confidence interval*—that is, a range within which the true parameter is located with *confidence level* probability. For example, a sample of several ^{14}C dates, belonging to a well-defined archaeological stratum (e.g., a destruction layer), may yield an uncalibrated confidence interval of 3000±30 years BP. A pre-existing ^{14}C calibration curve can then be utilized to derive a confidence interval of 1129–1275 BCE, within which the proper date is located, with a confidence level of 95%. For other types of samples, when no external probabilistic data akin to the calibration model exists, the sample itself can be used to estimate the confidence interval via procedures such as "bootstrap" (Efron 1979) or "jackknife" (Quenouille 1949).
- *P-value* estimation: This technique is based on contrasting two competing hypotheses. Commonly, hypothesis H_1 claims the "uniqueness" or "significance" of the sample. An *alternative* hypothesis H_0 usually claims that the results were obtained by mere chance. The p-value represents the probability of obtaining the observed sample under the H_0 ("null") hypothesis. Therefore, a low p-value represents high significance of H_1. For example, the sample may contain 75% Chalcolithic and 25% Late Byzantine pottery. H_1 would be "the site was larger during the Chalcolithic period than in the Late Byzantine period." Alternatively, H_0 would claim that the site had similar size during the two periods, and that the result was obtained at random. How "extraordinary" is the 75%–25% result? It depends on the actual amount of sherds of each type. If the sample consists of 3 Chalcolithic versus 1 Late Byzantine sherd, the p-value is 0.625, which is quite high, indicating an *insignificant* result. On the other hand, 30 Chalcolithic versus 10 Late Byzantine sherds yields a p-value of 0.002, which indicates *high significance* of H_1 hypothesis (again, when other assumptions are met).
- *Maximum likelihood* estimation: In certain situations, it is challenging to derive a statistical parameter directly. However, we can define a score indicating the "goodness" of this parameter in representing the statistical data at our disposal. Thus, in some sense, maximizing the score yields the "best" estimate of the parameter value (adhering to some underlying assumptions). For example, two types of Herodian oil lamps are found at the same stratum, with 5 lamps of Type A and 15 lamps of Type B. Assuming similar preservational characteristics, what was the original relative proportion of Type A lamp? Of course, it could have been 5% or 50% or even 99%, as any given assemblage is accidental. However, we can provide a statistical score quantifying the likelihood of each proportion. Such a procedure would yield 25% as the most likely result for the original Type A proportion. Although this is not always

the case, in our example, the estimation correlates with a common-sense intuition.

In many types of archaeological inquiry, the discussion is based on excavated finds. There either was an altar excavated within a specific stratum of a given site, or there wasn't one. But in other cases, the reasoning is based on statistical analysis of empirical data—for example, the date of an olive seed or proportions of various types or periods within a given ceramic assemblage. We believe that the purely descriptive approach is insufficient. With no statistical inference tools backing the empirical quantitative estimations, such data cannot be considered trustworthy.

Test Cases

Our joint endeavors with Prof. Finkelstein have produced several research threads. At the core of all of these studies lies a utilization of inferential statistics. Below, we provide several examples of our results and methods.

^{14}C Dating

Carbon-based dating is inherently a confidence interval estimation procedure. The method may be quite sensitive to the input data. Therefore, a strict data-handling protocol is necessary. Such a procedure is detailed in Finkelstein and Piasetzky 2010; what follows is a short summary.

- *The sources*: The data comes exclusively from strata with well-defined ceramic phases, in sequential horizons; see Finkelstein and Piasetzky 2015 for the importance of this point.
- *Selection of data for the model*: All available determinations from loci that are both safely assigned stratigraphically and well classified from the point of view of ceramic typology. Some of the strata that provided samples for radiocarbon dating had come to an end in heavy conflagrations that resulted in a thick collapse that had buried pottery vessels and other finds as well as clusters of charred grain seeds and olive pits. These cases provide the most secure provenance for radiocarbon dating. Due to the risk of "old wood effect," only short-lived samples are included (see Finkelstein and Piasetzky 2010 for methodological discussion on this topic). Only determinations that differ by more than 5 standard deviations from the weighted average of the other measurements in their group are excluded as outliers.
- *Exclusion of data by the model*: A Bayesian analysis using the OxCal program (Bronk Ramsey 1995, 2001) is performed. The individual agreement index calculated by OxCal checks how soundly the initial model agrees with each individual datum. In order to achieve a high level of agreement between the data and the model, the samples indicated as being extremely inconsistent with the model are removed until a high level of agreement is reached.

In addition to these careful data-cleansing steps, the resulting model contains *overlapping* confidence intervals for the different periods. Moreover, fig. 1 (adapted from Finkelstein and Piasetzky 2010) demonstrates not only the confidence intervals of each period but also confidence intervals for *transitions* between each two consecutive phases. This approach allows for some fine-grained, and at the same time quite cautious, statistically sound dating.

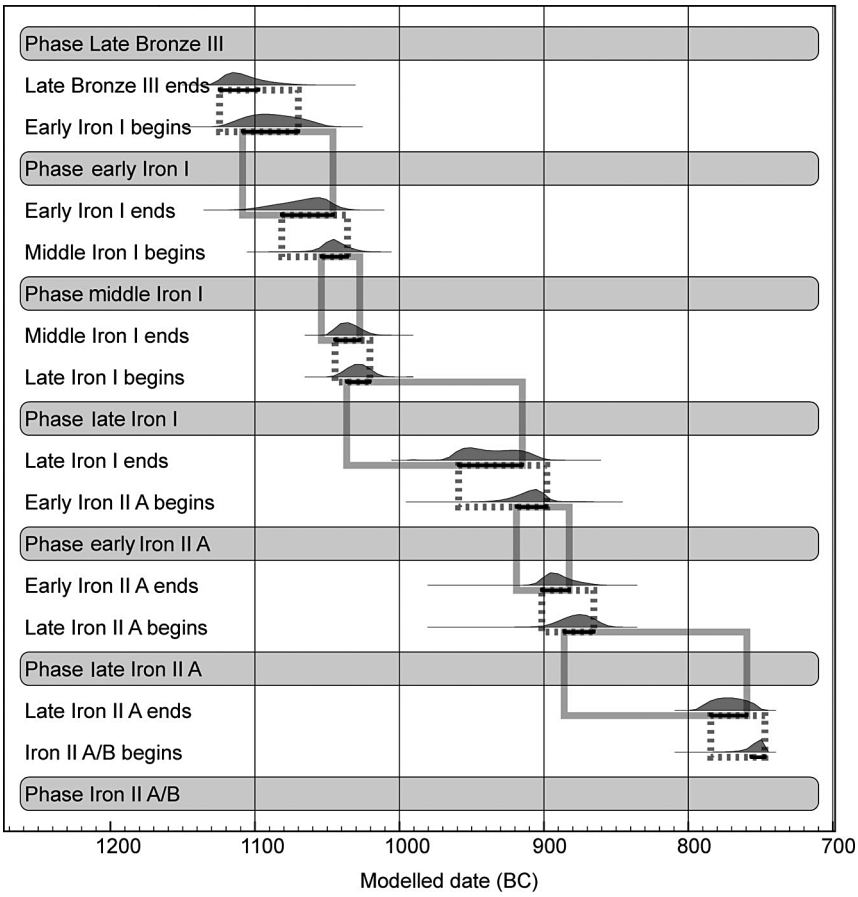

Fig. 1. A graphic depiction of confidence intervals for distinct chronological phases, beginning with Late Bronze III and ending with Iron II A/B. The figure shows not only the confidence intervals of each period but also confidence intervals for transitions between each two consecutive phases (adapted from Finkelstein and Piasetzky 2010).

Estimation of Number of Writers within a Corpus of Ostraca

In the paper Faigenbaum-Golovin et al. 2016, we focused on finding a minimal number of writers within the short-lived corpus of Iron Age inscriptions unearthed at the Arad fortress (Aharoni 1981). The issue was handled on a pair-wise basis. On this level, the question was: what is the probability that two given texts were written by the same author? Similar questions are frequently addressed by posing an *alternative* null hypothesis H_0 and attempting *to reject it*. In our case, for each pair of ostraca, the H_0 is: "both texts were written by the same author." At the first stage, various features were extracted from the characters of the inscriptions under examination (SIFT, Zernike coefficients, DCT, K_d-trees, projections, as well as L_1 and CMI distances. Regarding the last measure, specifically designed for ostraca, see Shaus et al. 2010, 2012a, 2012b). Subsequently, a statistical experiment testing the

Table 1. Comparison between different Arad inscriptions, adapted from (Faigenbaum-Golovin et al. 2016). A p-value ≤0.2, highlighted in gray, indicates rejection of "single-writer" hypothesis, hence accepting a "two-different-authors" alternative.

	Inscriptions' Content	1	2	3	5	7	8	16	17a	17b	18	21	24	31	38	39a	39b	40	111
1	Order to Eliashib, supply of provisions for the Kittiyim		0.64	0.50	0.91	0.30	0.64	0.51	0.98	0.78	0.53	0.24	0.003	0.10	0.27	0.41	0.06	0.23	v
2	Order to Eliashib, supply of provisions for the Kittiyim	0.64		1.00	1.00	0.72	1.00	0.39	0.85	0.78	0.31	0.75	0.79	0.06	0.38	0.98	0.70	0.11	0.96
3	Order to Eliashib mentioning Hananyahu, concerning provisions to Beer Sheba	0.50	1.00		0.23	0.06	0.55	0.36	1.00	0.77	0.27	0.94	0.72	0.16	0.61	0.96	0.84	0.22	0.79
5	Order to Eliashib, supply of provisions, probably for the Kittiyim	0.91	1.00	0.23		0.53	0.60	0.60	0.19	0.40	0.07	0.46	0.12	0.01	0.40	0.24	0.21	0.07	0.98
7	Order to Eliashib, supply of provisions for the Kittiyim	0.30	0.72	0.06	0.53		0.03	0.76	0.17	0.48	0.004	0.43	0.05	0.07	0.27	0.35	1.00	0.15	0.05
8	Order to Eliashib, supply of provisions for the Kittiyim	0.64	1.00	0.55	0.60	0.03		0.68	0.07	1.00	0.17	0.33	0.74	0.42	0.20	0.67	1.00	1.00	0.93
16	Letter to Eliashib from Hananyahu	0.51	0.39	0.36	0.60	0.76	0.68		0.33	1.00	0.03	0.80	0.13	0.38	0.38	0.41	0.40	0.72	0.68
17a	Order to Nahum to proceed to the house of Eliashib in order to collect provisions	0.98	0.85	1.00	0.19	0.17	0.07	0.33		1.00	0.92	0.36	0.13	0.41	1.00	0.68	1.00	0.17	0.68
17b	Note that Nahum provided provisions to the Kittiyim	0.78	0.78	0.77	0.40	0.48	1.00	1.00	1.00		1.00	0.35	0.40	0.47	1.00	1.00	0.33	0.20	0.40
18	Report to Eliashib from a subordinate fulfilling an order; mention of the Temple	0.53	0.31	0.27	0.07	0.004	0.17	0.03	0.92	1.00		3×10-4	0.02	0.20	0.32	0.94	0.86	0.04	0.73
21	Letter to Gedalyahu from a subordinate, Yehokal	0.24	0.75	0.94	0.46	0.43	0.33	0.80	0.36	0.35	3×10-4		0.35	0.04	0.23	0.71	0.21	0.31	0.90
24	A royal decree ordering the reinforcement of Ramat Negeb against Edom	0.003	0.79	0.72	0.12	0.05	0.74	0.13	0.13	0.40	0.02	0.35		0.01	0.05	0.73	0.38	0.002	0.92
31	List of names	0.10	0.06	0.16	0.01	0.07	0.42	0.38	0.41	0.47	0.20	0.04	0.01		0.33	0.16	0.11	0.35	0.57
38	List of names (inc. the son of Eliashib)	0.27	0.38	0.61	0.40	0.27	0.20	0.38	1.00	1.00	0.32	0.23	0.05	0.33		0.77	0.33	0.70	0.77
39a	List of names	0.41	0.98	0.96	0.24	0.35	0.67	0.41	0.68	1.00	0.94	0.71	0.73	0.16	0.77		1.00	0.04	0.75
39b	List of names	0.06	0.70	0.84	0.21	1.00	1.00	0.40	1.00	0.33	0.86	0.21	0.38	0.11	0.33	1.00		0.42	0.42
40	Gemaryahu & Nehemyahu report to Malkiyahu mentioning Edom and the king of Judah	0.23	0.11	0.22	0.07	0.15	1.00	0.72	0.17	0.20	0.04	0.31	0.002	0.35	0.70	0.04	0.42		0.67
111	Fragmentary, mentioning guard and horses	0.79	0.96	0.79	0.98	0.05	0.93	0.68	0.68	0.40	0.73	0.90	0.92	0.57	0.77	0.75	0.42	0.67	

plausibility of the null hypothesis was carried out for each letter independently (i.e., *alep*, *waw*, etc), and an appropriate p-value indicating the probability of obtaining an affirmative answer to H_0 was calculated. Upon obtaining several independent p-values (one for each letter), they were combined via the classic Fisher's method. Finally, the resulting "combined" p-value was considered. If the experiment's result was unlikely under the H_0 premise (p-value ≤ 0.2), we concluded that the two documents were written by two different individuals (i.e., we rejected the null hypothesis). On the other hand, if the occurrence of H_0 was probable (i.e., p-value > 0.2), we remained *agnostic*. Thus, in the latter case, we could not determine whether the two texts were in fact written by a single author or not.

A table summarizing the results of the experiment on the Arad corpus, with quite a few successful "hands" separations (i.e., rejected null hypotheses), can be seen at Table 1.

In the article Shaus and Turkel 2017, this methodology was advanced further with even more fine-grained "binary pixel patterns" features, providing an ample amount of additional experiments (per letter and per each feature!). This resulted in more significant results—similar to, but achieved independently from Faigenbaum-Golovin et al. 2016. The results of the experiment on the Arad corpus, again with many "hands" separations, can be observed in Table 2.

The outputs of both algorithms can be further assessed by analyzing the *overall* results, summarized by the corresponding tables. Initially, we can ask: how many pair-wise distinct authors can be found in the corpus (without taking the inscriptions' content into consideration)? It turns out that the answer is 4 in the case of Faigenbaum-Golovin et al. 2016 and 5 in case of Shaus and Turkel 2017. Moreover, in the former case, we observe six pair-wise distinct "quadruplets" of texts (namely, 7, 17a, 24, and 40; 5, 17a, 24, and 40; 7, 18, 24, and 40; 5, 18, 24, and 40; 7, 18, 24, and 31; 5, 18, 24, and 31), while in the latter case three pair-wise distinct "quintuplets" of inscriptions is demonstrated (1, 2, 18, 38, and 40; 1, 18, 24, 38, and 40; 5, 18, 24, 38, and 40). Now, on a meta-analysis level, we can ask: what is the probability for obtaining this number of quadruplets and quintuplets at random? Statistical simulations demonstrate that, in the former case, the probability (i.e., a "meta" p-value) is 2.17×10^{-4}, while in the latter, case it is 8×10^{-7}. Hence statistical significance of the results is very high in the former case, and is in fact much higher in the latter. Moreover, the results of these studies, with at least 4 unique writers in Faigenbaum-Golovin et al. 2016 and at least 5 unique writers in Shaus and Turkel 2017, independently confirm each other.

Raman-Based Image Acquisition Method

Our team inspected several research directions potentially allowing for image acquisition of ancient ostraca. Among the techniques studied were the multispectral imaging (see examples in Faigenbaum et al. 2012, 2014, 2015, Sober et al. 2014, Faigenbaum-Golovin et al. 2015a), as well as methods based on XRF (Nir-El et al. 2015) and Raman (Faigenbaum-Golovin et al. 2015b; Faigenbaum-Golovin et al. 2017) spectroscopies.

Table 2. Comparison between Various Arad Inscriptions, Adapted from (Shaus and Turkel 2017). A p-value ≤0.1, highlighted in gray, indicates rejection of "single writer" null hypothesis,

Text	1	2	3	5	7	8	16	17a	17b
1		0.000104451	0.794037056	0.997286098	0.835555244	0.999986905	0.145123589	0.999994862	0.345685406
2	0.000104451		0.999999997	0.230810507	0.23103925	0.999837107	0.782882802	0.999377805	0.121018637
3	0.794037056	0.999999997		0.99999999	1	0.999999995	1	1	0.999917956
5	0.997286098	0.230810507	0.99999999		0.999999996	0.999999979	0.829655242	0.955254622	0.093782225
7	0.835555244	0.23103925	1	0.999999996		0.999049432	0.974036343	0.938473936	0.543854905
8	0.999986905	0.999837107	0.999999995	0.999999979	0.999049432		0.996867196	0.999999995	0.26481398
16	0.145123589	0.782882802	1	0.829655242	0.974036343	0.996867196		0.914322562	0.989523149
17a	0.999994862	0.999377805	1	0.955254622	0.938473936	0.999999995	0.914322562		0.98838179
17b	0.345685406	0.121018637	0.999917956	0.093782225	0.543854905	0.26481398	0.989523149	0.98838179	
18	0.012774181	1.96989×10^{-08}	4.13858×10^{-07}	2.78436×10^{-17}	6.1151×10^{-25}	0.000160748	6.53819×10^{-37}	0.999824883	0.989432496
21	0.012535286	0.677687539	1	0.996839946	0.953218488	0.99999487	0.99999911	0.989761454	0.987354188
24	3.65925×10^{-12}	0.907018003	1	4.38907×10^{-10}	0.01448862	0.995623889	0.050144018	0.916602116	0.397038359
31	0.016073174	0.460683411	0.51882444	0.026192004	0.15146191	0.490501387	1.82513×10^{-08}	0.999934002	0.997936364
38	0.078182536	1.45921×10^{-07}	0.960818768	0.026797057	0.651369422	0.028988689	0.017541713	0.971750814	0.994674672
39a	0.038794364	0.999999999	1	0.942902756	0.973278889	0.999933169	0.999999999	0.999957681	0.690349079
39b	0.308694925	0.004777329	0.999999219	0.867249786	0.381286518	0.610834495	0.752972486	0.999885898	0.989449919
40	2.97103×10^{-06}	0.011757086	0.983116915	0.040887838	0.999602961	0.999979451	0.986154478	0.750470353	0.469371025
111	0.000119896	0.918537018	1	2.1339×10^{-05}	0.072003619	0.818431322	0.999998263	0.999999995	0.976381275

Our initial Raman spectroscopy experiments showed a clear distinction between clay and ink spectra, which was utilized to construct a macroscale mapping device. The spectral differences recorded by the device and detected by specially engineered algorithms allowed for production of new automated facsimiles (black and white images) of the inscriptions. Our method circumvented the preparatory ink composition analysis (common in Raman spectroscopy), allowing for a straightforward detection of indicative Raman lines (wavelengths). Utilizing these lines, the most legible facsimiles were obtained.

[Table 2, cont.] hence accepting a "two different authors" alternative. Note the typically high statistical significance, i.e., very low p-values upon rejection of null hypothesis.

18	21	24	31	38	39a	39b	40	111
0.012774181	0.012535286	3.65925×10^{-12}	0.016073174	0.078182536	0.038794364	0.308694925	2.97103×10^{-06}	0.000119896
1.96989×10^{-08}	0.677687539	0.907018003	0.460683411	1.45921×10^{-07}	0.999999999	0.004777329	0.011757086	0.918537018
4.13858×10^{-07}	1	1	0.51882444	0.960818768	1	0.999999219	0.983116915	1
2.78436×10^{-17}	0.996839946	4.38907×10^{-10}	0.026192004	0.026797057	0.942902756	0.867249786	0.040887838	2.1339×10^{-05}
6.1151×10^{-25}	0.953218488	0.01448862	0.15146191	0.651369422	0.973278889	0.381286518	0.999602961	0.072003619
0.000160748	0.99999487	0.995623889	0.490501387	0.028988689	0.999933169	0.610834495	0.999979451	0.818431322
6.53819×10^{-37}	0.99999911	0.050144018	1.82513×10^{-08}	0.017541713	0.999999999	0.752972486	0.986154478	0.999998263
0.999824883	0.989761454	0.916602116	0.999934002	0.971750814	0.999957681	0.999885898	0.750470353	0.999999995
0.989432496	0.987354188	0.397038359	0.997936364	0.994674672	0.690349079	0.989449919	0.469371025	0.976381275
	1.74207×10^{-37}	2.85314×10^{-23}	0.872302919	6.42961×10^{-05}	0.000351619	3.9727×10^{-05}	4.06656×10^{-14}	0.016251091
1.74207×10^{-37}		0.004905952	0.002746089	0.122737346	0.999999846	0.982566541	0.957807456	0.999809072
2.85314×10^{-23}	0.004905952		2.93422×10^{-09}	0.009250254	0.999999991	0.405106241	6.4106×10^{-12}	0.684834031
0.872302919	0.002746089	2.93422×10^{-09}		0.856161514	0.99999999	0.999884872	1.70082×10^{-09}	0.772576882
6.42961×10^{-05}	0.122737346	0.009250254	0.856161514		0.600342228	0.112003203	0.067411667	0.09199445
0.000351619	0.999999846	0.999999991	0.99999999	0.600342228		0.999985794	0.046450883	1
3.9727×10^{-05}	0.982566541	0.405106241	0.999884872	0.112003203	0.999985794		0.839063568	0.90756289
4.06656×10^{-14}	0.957807456	6.4106×10^{-12}	1.70082×10^{-09}	0.067411667	0.046450883	0.839063568		0.113447052
0.016251091	0.999809072	0.684834031	0.772576882	0.09199445	1	0.90756289	0.113447052	

The method was tested on an Edomite ostracon from Ḥorvat ʿUza (Beit-Arieh 2007). The scans were performed on a character level. In fig. 2, a mapping result of one character, as well as facsimiles created after various postprocessing steps, can be seen.

Given the intricacies of signal acquisition and processing procedures, a legitimate question still remains: could the obtained binary mapping have been obtained "at random"? This brings us back to the concept of p-value. Indeed, no type of spectral noise can produce a meaningful Raman mapping. The resulting binarization

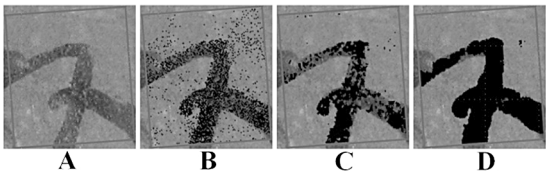

Fig 2. Raman mapping of a single character (adapted from Faigenbaum-Golovin et al. 2015b). (A) photograph of the mapping area; (B) mapping result for Raman line ~1460 cm^{-1} overlaid; (C) mapping result after post-processing, configuration I: median filter, window of 7×7 pixels and unifying 5 neighboring images with respect to the selected wavelength; (D) mapping result after post-processing, configuration II: percentile filter, window of 9×9 pixels, percentile=20%.

implies that the results possess an extremely high statistical significance. Namely, the probability of obtaining a binarization with similar (or better) correlation to the ostracon at random (i.e., the p-value) is estimated to be less than 10^{-30}! In fact, the probability of obtaining a "random" binary image depicting a legitimate written ligature is of similar order of magnitude to that of a monkey accidentally typing the introduction to this article. In other words, the binarization results are by all means significant.

Conclusions (versus "Conclusions")

Archaeology is a respected and well-established branch of learning. In recent years, it has slowly absorbed new and effective techniques from other disciplines, mainly from the life and exact sciences. This provides a suitable opportunity for the reassessment of the desired focus of archaeological research. Indeed, it is our view that, instead of being utilized as auxiliary devices, exact and inferential approaches ought to be essential in archaeological study. Research questions should allow for precise data collection and employment of inferential statistical analysis.

Admittedly, some archaeological research is performed in this vein even today. A random sample of relatively recent papers (Drennan and Peterson 2004, Markofsky 2014, de Pablo and Barton 2015) shows that some scholars are indeed aware of the possibilities that statistical inference has to offer. However, this is an exception, rather than the norm. Indeed, even in the "Proceedings of the 43rd Annual Conference on Computer Applications and *Quantitative Methods in Archaeology*" (Campana et al. 2015), a venue that one would expect to be suitable for inferential statistics, out of 116 papers in 13 different sessions, only two (1.72%) mentioned "significance interval"; two (1.72%) mentioned "p-value" (one of them claimed a p-value of 0, which is impossible in statistics, since however unlikely, any results could have been obtained accidentally); and only one (0.86%) mentioned "significance" in its statistical meaning. In other words, statistical inference was significantly underrepresented (in both colloquial and statistical meanings), even in this apparently appropriate event.

Nevertheless, it is our conviction that this deficiency is about to, and ought to, be changed. If archaeology as a discipline wishes to proceed further, it needs to be able to quantify the validity of its conclusions and to center its investigation around this point. We envisage future inquiries, including excavations, as tightly controlled sampling mechanisms, designed specifically to provide diverse, unbiased, and representative samples. The end goal would be the possibility of conducting statistical tests supporting (or refuting) the deductions.

Thus, although it may seem that the current paper deals with complicated statistical procedures and their application to archaeological research, in fact it is all about the conclusions. What's at stake here is *the possibility to quantify the validity of archaeological conclusions*. Preferably, this quantification will pertain to *the probability of obtaining an incorrect conclusion*. If such a possibility is lacking, the "conclusions" of such research might be unwarranted. Alternatively, the existence of such a procedure may represent the core requirement of modern scientific archaeology.

Acknowledgments

The research reported here received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no. 229418, and by an Early Israel grant (New Horizons project), Tel Aviv University. This study was also supported by a generous donation from Jacques Chahine, made through the French Friends of Tel Aviv University. Arie Shaus is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. The kind assistance of Shirly Ben-Dor Evian, Sivan Einhorn, Noa Evron and Myrna Pollak is greatly appreciated. Ostraca and facsimile images are courtesy of the Institute of Archaeology, Tel Aviv University, and of the Israel Antiquities Authority.

References

Aharoni, Y. 1981. *Arad Inscriptions*, Jerusalem.

Arnold III, P. J. and Wilkens, B. S. 2001. On the Vanpools' "Scientific" Postprocessualism. *American Antiquity* 66.2: 361–66.

Beit-Arieh I. 2007. *Ḥorvat ʿUza and Ḥorvat Radum: Two Fortresses in the Biblical Negev* (Monograph Series of the Institute of Archaeology 25). Tel Aviv. Claire Yass Publications in Archaeology.

Binford L. 1962. Archaeology as Anthropology. *American Antiquity* 28.2: 217–25.

Bronk Ramsey, C. 1995. Radiocarbon Calibration and Analysis of Stratigraphy: The OxCal Program. *Radiocarbon* 37: 425–30.

Bronk Ramsey, C. 2001. Development of the Radiocarbon Program OxCal. *Radiocarbon* 43: 355–63.

Campana, S., Scopigno, R., Carpentiero, G. and Cirillo, M. 2015. *Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2015)*. Oxford.

de Pablo, J. F. L. and Barton, C. M., 2015. Bayesian Estimation Dating of Lithic Surface Collections. *Journal of Archaeological Method and Theory*, 22.2: 559–83.

Drennan, R. D. and Peterson, C. E. 2004. Comparing Archaeological Settlement Systems with Rank-Size Graphs: A Measure of Shape and Statistical Confidence. *Journal of Archaeological Science* 31.5: 533–49.

Drower, M. S. 1995. *Flinders Petrie: A Life in Archaeology*. Madison.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7.1: 1–26.

Faigenbaum, S., Sober, B., Shaus, A., Moinester, M., Piasetzky, E., Bearman, G., Cordonsky, M. and Finkelstein, I. 2012. Multispectral Images of Ostraca: Acquisition and Analysis, *Journal of Archaeological Science* 39.12: 3581–90.

Faigenbaum, S., Sober, B., Finkelstein, I., Moinester, M., Piasetzky, E., Shaus, A. and Cordonsky, M. 2014. Multispectral Imaging of Two Hieratic Inscriptions from Qubur El-Walaydah. *Ägypten und Levante* 24: 349–53.

Faigenbaum, S., Sober, B., Moinester, M., Piasetzky, E. and Bearman, G. 2015. Multispectral Imaging of Tel Malhata Ostraca. In: Beit-Arieh, I., ed. *Tel Malhata: A Central City in the Biblical Negev* (Monograph Series of the Institute of Archaeology 32). Tel Aviv: 510–13.

Faigenbaum-Golovin, S., Rollston, C. A., Piasetzky, E., Sober, B. and Finkelstein, I. 2015a. The Ophel (Jerusalem) Ostracon in Light of New Multispectral Images. *Semitica* 57: 113–37.

Faigenbaum-Golovin, S., Shaus, A., Sober, B., Finkelstein, I., Levin, D., Moinester, M., Piasetzky, E. and Turkel, E. 2015b. Computerized Paleographic Investigation of Hebrew Iron Age Ostraca. *Radiocarbon* 57.2: 317–25.

Faigenbaum-Golovin, S., Shaus, A., Sober, B., Levin, D., Na'aman, N., Sass, B., Turkel, E., Piasetzky, E., Finkelstein, I. 2016. Algorithmic Handwriting Analysis of Judah's Military Correspondence Sheds Light on Composition of Biblical Texts. *Proceedings of the National Academy of Sciences* 113.17: 4664–69.

Faigenbaum-Golovin, S., Mendel-Geberovich, A., Shaus, A., Sober, B., Cordonsky, M., Levin, D., Moinester, M., Sass, B., Turkel, E., Piasetzky, E., and Finkelstein, I. 2017. Multispectral Imaging Reveals Biblical-Period Inscription Unnoticed for Half a Century. *PLOS ONE* 12.6: e0178400.

Finkelstein, I., and Piasetzky, E. 2010. Radiocarbon Dating the Iron Age in the Levant: A Bayesian Model for Six Ceramic Phases and Six Transitions. *Antiquity* 84: 374–85.

Finkelstein, I., Boaretto, E., Ben Dor Evian, S., Cabanes, D., Cabanes, M., Eliyahu, A., Faigenbaum, S., Gadot, Y., Langgut, D., Martin, M., Meiri, M., Namdar, D., Sapir-Hen, L., Shahack-Gross, R., Shaus, A., Sober, B., Tofollo, M., Yahalom-Mack, N., Zapassky, L. and Weiner, S. 2012. Reconstructing Ancient Israel: Integrating Macro- and Micro-archaeology. *Hebrew Bible and Ancient Israel* 1: 133–50.

Finkelstein, I., Weiner, S. and Boaretto, E. 2015. Preface—The Iron Age in Israel: The Exact and Life Sciences Perspectives. *Radiocarbon* 57.2: 197–206.

Finkelstein, I. and Piasetzky, E., 2015. Radiocarbon Dating Khirbet Qeiyafa and the Iron I–IIA Phases in the Shephelah: Methodological Comments and a Bayesian Model. *Radiocarbon* 57.5: 891–907.

Flannery, K. V. 1982. The Golden Marshalltown: A Parable for the Archaeology of the 1980s. *American Anthropologist* 84: 265–78.

Frisch, R. 1933. Editor's Note. *Econometrica* 1.1: 1–4.

Hutson, S.R 2001. Synergy through Disunity, Science as Social Practice: Comments on Vanpool and Vanpool. *American Antiquity* 66.2: 349–360.

Markofsky, S. 2014. When Survey Goes East: Field Survey Methodologies and Analytical Frameworks in a Central Asian Context. *Journal of Archaeological Method and Theory* 21.4: 697–723.

Nir-El, Y., Goren, Y., Piasetzky, E., Moinester, M. and Sober, B. 2015. X-ray Fluorescence (XRF) Measurements of Red Ink on a Tel Malhata Ostracon. In: Beit-Arieh, I., ed. *Tel Malhata: A Central City in the Biblical Negev* (Monograph Series of the Institute of Archaeology 32). Tel Aviv: 507–9.

Popper, K. R. 1960. *The Poverty of Historicism* (2nd ed.), London: 142.

Quenouille, M. H. 1949. Problems in Plane Sampling. *The Annals of Mathematical Statistics* 20.3: 355–75.

Shaus, A., Finkelstein, I. and Piasetzky, E. 2010. Bypassing the Eye of the Beholder: Automated Ostraca Facsimile Evaluation. *Maarav* 17: 7–20.

Shaus, A., Turkel, E. and Piasetzky, E. 2012a. Quality Evaluation of Facsimiles of Hebrew First Temple Period Inscriptions. *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS 2012)*: 170–74.

Shaus, A., Turkel, E. and Piasetzky, E. 2012b. Binarization of First Temple Period Inscriptions: Performance of Existing Algorithms and a New Registration Based Scheme. *Proceed-*

ings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012): 645–50.

Shaus, A. and Turkel, E. 2017. Writer Identification in Modern and Historical Documents via Binary Pixel Patterns, Kolmogorov-Smirnov Test and Fisher's Method. *Journal of Imaging Science and Technology* 61.1: 010404-1-010404-9.

Sober, B., Faigenbaum, S., Beit-Arieh, I., Finkelstein, I., Moinester, M., Piasetzky, E. and Shaus, A. 2014. Multispectral Imaging as a Tool for Enhancing the Reading of Ostraca. *Palestine Exploration Quarterly* 146: 185–97.

VanPool C. S. and VanPool T. L 1999. The Scientific Nature of Postprocessualism. *American Antiquity* 64.1: 33–53.

VanPool T. L and VanPool C. S. 2001. Postprocessualism and the Nature of Science: A Response to Comments by Hutson and Arnold and Wilkens. *American Antiquity* 66.2: 367–75.