Proceedings of the International Workshop on the Scientific approach to the Acheiropoietos Images, ENEA Frascati, Italy, 4-6 May 2010

A critical review of the radiocarbon dating of the Shroud of Turin. ANOVA - a useful method to evaluate sets of high precision AMS radiocarbon measurements

Remi Van Haelst

Kerkstraat 68 Bus 4 2060 Antwerp Belgium. remi.vanhaelst@scarlet.be

Abstract

A review of the radiocarbon literature illustrates limitations the AMS method shows when dating bones and all kind of living plants like flax, mainly composed out of cellulose. A number of old and recent radiocarbon dating results, made on linen and wood, are compared.

Keywords: AMS Statistics, Chi² test, IEM-EEM, ANOVA.

1. INTRODUCTION

In 1989, Nature [1] published the report on the radiocarbon dating of the Shroud of Turin, by the laboratories of Oxford, Arizona and Zurich. Claimed was of mediaeval date for the Shroud with a least 95% confidence.

I then published a small booklet [2], with a complete statistical analysis, including Chi², IEM-EEM and a small ANOVA [3] tests; showing that the claimed 95 % confidence was not supported by a statistical test. Today ANOVA is accepted by NIST [4].

When I conducted some heating experiments on inducing ^{14}C enrichment, I received an official dating report from the Oxford radiocarbon laboratory. I was surprised to read the following caveat:

"One should bear in mind that these measurements have been made on organic material and that this cannot be regarded as a guarantee of the article date of manufacture. It should be noted that the undetected presence of any contaminant may affect any radiocarbon result."

A caveat in contrast with the more stringent requirements imposed for industrial laboratories. In that context, a precision in part per million is mandatory.

2. AMS MEASUREMENTS

Until 1977, radiocarbon measurements were made by counting the number of ^{14}C decays over a long period. The development of AMS with real ^{14}C isotope counting was a revolution. One became able to almost completely separate nitrogen and carbon isotopes. In 1986, a Zurich AMS test run with 14 standard samples, showed a counting error of 0.3% and an overall error of ~ 1 %. The conclusion of that study was: "One should improve the count rate and reduce contamination by at least a factor 2" [5]. So, the need to improve the precision of AMS measurements has been known for some time.

Scott, analyzing the 1990 International Collaborative Programme [6] concluded: "*It seems reasonable to consider that a laboratory performs adequately if it has no systematic bias and assesses its internal and external variability adequately. IEM & EEM should not significantly different from 1. In total, only 15 out of 58 laboratories did meet these 3 basic criteria*".

Today in its website, the AMS laboratory of the University of Arizona at Tucson [7] claims an error as low as 0.2%, a high level of precision in RC year terms. They note that when results are in doubt, measurements are repeated and the AMS equipment tuned or even shut-down for repair.

This level of precision and accuracy has changed over time. In 1990-1992, Arizona obtained ~ 85% correct measurements. By 2000, they improved to about 92% but this still means nearly one failure out of ten measurements. May one wonder about the failure rate in 1988?

It should also be noted the need for a statistical analysis of the ^{14}C count, due to some unpredictable spontaneous reactions:

$$^{14}N + n = {}^{14}C + p \;\; \& \;\; {}^{14}C \Rightarrow {}^{14}N + b + 160 \text{ keV}$$

Analysing char and root fractions from grain and pollen samples, NIST researchers [8] noted significant differences in the ^{14}C content of different fractions taken from the same sample. Analysis of SRM 1649a NIST reference material showed an elemental ^{14}C char/soot ratio of 2.75. The biomass is about 38% and contains a mixture of about 13% aromatic components. Because such a high biomass carbon fraction is very important, there must be a significant missing carbon component in this material.

Most important, it was noted that cellulose (such as the linen of the Shroud) is an excellent candidate for easy contamination!

Recent "molecular analysis" of individual amino acids from crude collagen and gelatine fractions from the Dent Mammoth [8], shows ^{14}C counts between 4000 and 2500 (8000 ~ 11000 BP).

In *Radiocarbon* n° 40 (1996), the same author noted: "*In the nearby future modern ^{14}C techniques will eventually lead to the application of a **real** isotopic mass balance, using actual **true** ^{14}C counting*".

As recommended by Polach [9], ^{14}C count values are more appropriate in analysis than using "RC ages" which are log-normally distributed. An example of this effect is as follows.

A RC age of 795 ± 65 years represents an uncertainty range of 65/795 or ± 8.2%.
The equivalent ^{14}C data produces a count of 8629 ± 68 which represents an error of 68/8267 = 0.82%.

Likewise,

795 + 65 = 860; 860/8267 = 0.104015; exp(0.104015) = 1.1096; 9500/1.1096 = 8561.5 ^{14}C count
795 - 65 = 730; 730/8267 = 0.088292; exp(0.088292) = 1.0923; 9500/1.0923 = 8697.2 ^{14}C count
Mean date = 8629 +/- 68.

From this it is apparent that relatively small changes in ^{14}C count can translate into large changes in reported radiocarbon dates.

3. COMPARISON OF A SET OF DATES, REPORTED BY POLACH AND THE DATES REPORTED FOR THE SHROUD

To illustrate how best to evaluate radiocarbon data, Wilson and Ward [9] used data for three independent measures of a single piece of wood, given by Polach [9].

From this we can conclude that there is no evidence to reject the null hypothesis the three samples observations are consistent.

Using these three samples with a calculated χ^2 value is a useful way to compare the *Nature* [1] data reported in Table 2 for the Shroud. Here we apply the same methodology as above to evaluate the hypothesis that the measurements are consistent.

As showed previously, one should correct Table 2 as follows (see the full analysis on www.shroud.com, paper by Van Haelst):

Comparison between the Polach samples and data given in Nature.

Polach	**Nature Table 2**	**Nature following Table 1**
Sample a 4330 ± 190	Arizona 646 ± 31	Arizona: 646 ± 17
Sample b 4560 ± 210	Oxford 750 ± 30	Oxford: 749 ± 31
Sample c 4940 ± 300	Zurich 676 ± 24	Zurich: 676 ±24
Mean: 4525 ± 128	Mean 689 ±16	Mean 672 ± 13
Chi² : 2.99 < 5.99	Chi² 6.35 > 5.99	Chi² 8.56 > 5.99
p = 0.24	p = 0.042	p = 0.012

Thus, in both the published and corrected cases, there is no reason to accept the null hypothesis that the observations are consistent and provide 95% confidence. However, for the Shroud measurements the radiocarbon researchers rejected this conclusion. One laboratory even questioned the statistical method used by the British Museum. According to Prof. Ramsey, Director of Oxford RC Laboratory, the measurements for the Shroud obtained in 1988 were within the acceptable error range of the AMS facilities of that time. He noted that he did not wish to spend his time recalculating data statistics [10].

Dr. Hedges (Oxford) and Prof. Jull (Arizona) agreed that there is indeed a "small" statistical problem, the Oxford dates being different from the two other laboratories [10, 11]. Unfortunately, none of them answered the question: "How did you obtain the claimed 95% confidence?"

Another recent example of a possible erroneous radiocarbon result is the dating of the "Seamless Cloth"

[12], the type of garment mentioned in the Gospel of John (19:23). The cloth kept in Argenteuil, thought to be the Seamless Cloth, was twice radiocarbon dated by Gif-sur-Yvette and later in a totally blind evaluation by ETH Zurich. See below for the dating results.

Seamless Cloth dating results:		
Gif A 40100: 1450 ± 40	Gif A 40101: 1510 ± 40	ETH 30402: 1260 ± 40
Error weighted Mean: 1407 ± 23	χ^2 : 21.2917 > 5.66	p-value =0.0000

Thus, the hypothesis that the measurements are homogeneous and the means equal is rejected and the dating results are shown to be not conclusive.

Also the analytic details related to the Acid-Alkali-Acid cleaning are noticeable:

	Carbon	Oxygen	Aluminium	Sulphur	Calcium	Iron
Before	56	43	3	9	31	2
After	54	29	15	14	10	0

(height of the peaks in mm on the graphs.)

A loss in Carbon, Oxygen, Calcium and Iron. A gain in Aluminium and Sulphur.
A loss of about 1/3 in weight, probably indicating some heavy contamination.

It is clear from these two examples that there are apparent difficulties in reliably dating old fabric using standard radiocarbon dating precleaning techniques.

Today in AMS single run, one measures with repetition between 6 ~ 20 pure carbon targets prepared from the same sample, together with a number of standard and blank samples. The pure carbon is mixed with a graphite carrier. These pellets (targets) are placed on a turning wheel, to be measured one after another using sophisticated AMS equipment. Measured are also a number of standard samples and blanks. The targets are bombarded with high energy beams. Separation of ^{12}C, ^{13}C and ^{14}C isotopes is almost complete while care is taken to avoid crater formation in the targets.

Note that ^{14}C is counted, while the other isotopes are "frequency current" measurements. In Arizona the laboratory uses "coulomb/second" measurements. The measured ratio $^{14}C/^{13}C$ is about 18 times the natural ratio [7]. In practice, AMS measurements still are of variable precision. Therefore one needs corrections, taking in account a possible instrumental "drift".

Each laboratory uses a specific method to correct variable counts, taking into account the correction factors for the measured ratio $^{14}C/^{13}C$ and ^{14}C count and as a result, for this reason, "raw" count data cannot be used in statistical analysis. Several runs such as this are made to create a set of measurements with their standard errors. For instance, for the 1988 Shroud dating, Arizona made eight independent runs [10]. Each single AMS measurement is the combination of at least 6 observations per run.

Applying a classic analysis of variance (ANOVA) taking in account only the counted ^{14}C particles, allows to determine whether the measurement differences noted are due to chance or to the fact that the differences between the runs are indeed too large. By chance alone, the F statistic should be ~ 1.00. Errors are assumed to be due to chance or to experimental uncertainty.

4. ANOVA

In 1986, the British Museum applied an Analysis of Variance on the 12 individual measurements supplied by the laboratories, to determine the t_d value for 2 - 9 degrees of freedom [1]. They found that the errors based on the scatter should be multiplied by a factor 2.56 to more appropriately represent the variability in the data.

In the English version of a small booklet published in 1989 [2] I already employed ANOVA. Analysing the 12 mean data in Table 2 of the *Nature* paper [1], I concluded: "*The calculated F value 4.7 is larger than 4.2, the critical F value for 2-9 degrees of freedom*." With results like these, one should not draw any conclusions but ask for more and better measurements. Further, other researchers have also used ANOVA to analyse Table 2 and came to the same conclusion [13].

The accuracy of the ANOVA method can be impacted by differing numbers of measurements per group, large deviations from the normal distribution and inequalities in the variances of each of the groups being evaluated. Being sure these factors are accounted for, ANOVA provides a useful means of evaluating comparative measurements.

Using ^{14}C count means much tedious calculation work, but is readily made manageable by using an Excel worksheet or by using any of the modern commercially-available statistical packages.

In this study the laboratory data given in Table 1 will be analysed by ANOVA, taking into account the observation

that each single date is the results of multiple measurements. The errors based on quoted errors and in percent are used.

5. MODELLING

Let suppose there are three runs, each counting 10 ***standard*** samples (targets) with a number of blanks, measured under the same conditions, in the same AMS machine.

To simplify calculations, we assume that the exact number of ^{14}C counts for each standard sample totals 30000 with the measurements normally distributed ± 0.3% around the mean for each run. In our example, the total number of ^{14}C count is equal to 10,030 + 10,000 + 9,970 = 30,000.

With the above assumptions we observe that:

Run A	**Run B**	**Run C**
987	984	981
993	990	987
997	994	991
999	996	993
1002	999	996
1004	1001	998
1007	1004	1001
1009	1006	1003
1013	1010	1007
1019	1016	1013

We then use one-way ANOVA to evaluate the null hypothesis that the mean value of each of the runs is equal.

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*	S.D.
Run A	10	10030	1003	718	8.93
Run B	10	10000	1000	718	8.93
Run C	10	9970	997	718	8.93

ANOVA

Source of Variation	***SS***	***Df***	***MS***	***F***	***P-value***
Between Runs	180	2	90	1.02297	**0.3730**
Residual error	2375.44	27	87.98		
Total	2555.44	29			

Conclusion: the hypothesis that the mean value of each of the runs is the same is accepted.

A practical example

In the past, I received a breakout of the original measurements provided by the University of Arizona's Tucson radiocarbon dating facility to the British Museum as a part of the 1988 radiocarbon dating experiment [10]. These data (originally reported laboratory measurements) are:

1988 Radiocarbon dating experiment: original measurements with quoted error of measurement at 1 σ level.
Shroud of Turin samples

Laboratory	Measurement (RCYBP)	Error
	606	**+/- 41**
	574	**+/- 45**
	753	**+/- 51**
Arizona	**632**	**+/- 49**
	676	**+/- 59**
	540	**+/- 57**
	701	**+/- 47**
	701	**+/- 47**
	733	**+/- 61**
	722	**+/- 65**
Zurich	**635**	**+/- 57**
	639	**+/- 45**
	679	**+/- 51**
	795	**+/- 65**
Oxford	**730**	**+/- 45**
	745	**+/- 55**

The Arizona data were combined into four measurements and those measurements are the dates reported in the *Nature* paper [1]:

Original data:			Combined: (Van Haelst Acts CIELT Rome 1993 p 216)
Session A	606 ± 41	574-+45	591± 30
Session B	753 ± 51	632-+49	690 ± 35
Session C	540 ± 57	676-+59	606 ± 41
Session D	701 ± 47	701-+47	701 ± 47
Mean	646 ± 17		646 ± 17 (*Nature:* 647 ± 31) [1]

Unfortunately, the *Nature* paper never mentioned the combination of the eight observations into four observations and, as a result, the statistical analysis reported was somewhat misleading.

Because no information was provided by the laboratories, I was obliged to recalculate the number of ^{14}C atoms detected. I used this calculation to simulate a distribution of observations that make up each of the runs and test the hypothesis that the runs means are the same.

Let us assume the following characteristics: counting error = 0.3%, with eight runs each using 10 targets as shown in the following example:

Run A	Run B	Run C	Normally Distributed Run D	Run E	Run F	Run G	Run H
8539	8592	8591	8617	8664	8691	8725	8750
8589	8642	8641	8668	8714	8742	8776	8802
8618	8672	8671	8698	8745	8773	8806	8832
8642	8696	8695	8722	8769	8797	8831	8857
8664	8718	8717	8744	8791	8818	8852	8878

8684	8738	8737	8764	8811	8840	8874	8900
8706	8760	8759	8786	8833	8861	8895	8921
8730	8784	8783	8810	8857	8885	8920	8946
8759	8814	8813	8840	8888	8916	8950	8976
8809	8864	8863	8891	8938	8967	9001	9028

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Run A	10	86740	8674	6619.368
Run B	10	87280	8728	6702.043
Run C	10	87270	8727	6700.507
Run D	10	87540	8754	6742.032
Run E	10	88010	8801	6814.622
Run F	10	88290	8829	6858.051
Run G	10	88630	8863	6910.973
Run H	10	88890	8889	6951.580

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*
Between Runs	389,588.8	7	55,655.54	8.1998	0.0000
Residual error	488,692.6	72	6,787.40		
Total	878,281.3	79			

We reject the hypothesis that the means of each of these runs is equal and accept the hypothesis that one or more of the means is statistically different.

Detection of possible outliers using the method given by Burr [7]

Ratio: Count A/((Total – A)/7) = 8899/((70273 – 8899)/7) = 1.015
The same calculations for B, C, D, E, F, G and H.
All counts about 2σ away from the mean may be possible outliers.
Sum: 8899 8863 8829 8801 8754 8727 8727 8673 = 70273/8= 8784
Ratio 1.015 1.01 1.006 1.002 0.996 0.993 0.993 0.986 = 8.001/8 = 1.00125

The dates 8899 (= 540 yr) and 8673 (= 753 yr) are borderline results.

Interestingly, Christen, [15] analysing the Shroud data as given in [1] and using Bayesian statistics, came to the same conclusion: the dates 591 (Arizona) and 795 (Oxford) are possible outliers.

It should be noted that by simply discarding one outlier the Shroud data are more consistent.

Applying the IEM-EEM criteria, as proposed by Scott, leads to the same conclusion [6].

For Oxford and Arizona the External Error Multipliers are 1.45 and 1.5

ANOVA Analysis of the Shroud RC data, reconverted to ^{14}C count, based on the quoted errors.

	Targets	DF	Between	Residual	F ratio	97.5%	
Ox	18	2-15	13900/2 = 6950	46459/15 = 3097	6950/3097	= 2.2 < 5 =	OK
Ar	24	3-20	57492/3 = 19164	24563/20 =1227	19164/1227	= 15.6 > 4 =	FAILS
Zu	30	4-25	49603/4 = 12401	73815/25 = 2953	12401/2953	= 4.2 > 3.5 =	????
Mean	72	11-60	247390/11 = 22490	160312/60 = 2672	22490/2672	= 8.4 > 2.3	FAILS

Enlarging the errors for Zurich and Arizona, in order to obtain the critical F values, is not sufficient to obtain the critical F value for the combined 12 data.

6. CONCLUSION

The radiocarbon dating of cellulose-based textiles need to be approached very carefully since textiles appear to present experimental limitations which can result in non-homogeneous measurements.

Concerning the Shroud dating, the Arizona F value is out of range and should not be used in further calculations and certainly not in drawing conclusions supporting a 95% confidence.

The Zurich F value is a borderline case. In theory, the combination of 12 data is meaningless.

As stated by Burr, et. al. [7], one should verify the tuning the equipment and the effectiveness of the AAA cleaning methods before drawing any conclusions.

The different data for $-\delta^{13}C$: Oxford: 0.027, Arizona 0.025, Zurich 0.0251, given in Table 1 of Nature [1] indicate a further need to examine the homogeneity and the chemical composition of the twelve sub-samples.

ACKNOWLEDGMENTS

I like to thank: the Referees for their competent remarks and constructive suggestions; Bryan Walsh for his precious help and technical assistance and Diana Fulbright for inviting me to present this paper to the IWSAI Conference Frascati 2010.

NOTES AND REFERENCES

1. Damon et al., Nature, **337**, 611-615 (1989).

2. R. Van Haelst "*Radiocarbon dating the Shroud of Turin.*" Privately released. (1989).

3. NIST Technical note. B Taylor &C. Kuyatt (1994)

4. Perry's Chemical Engineering Handbook (Fourth Edition). The practical example of the F-test given on pages 2.72-74 was used to test the Excel programme.

5. Suter et al. Nuclear Instruments and Methods. **223**. (1986).

6. Scott et al. Antiquity Volume 64 (June 1990) & International Collaborative Programme 'Trondheim Radiocarbon Conference.

7. Burr et al. Nuclear Instruments and Methods B **259** 149-153 (2007).

8. Lloyd, Currie: "*The Remarkable Metrological History of Radiocarbon Dating*". (Part II in NIST volume 109 N° 2 2004) & (Radiocarbon 40 1998). See also Czechoslovak Journal of Physics **53**, cited by Lloyd.

9. Wilson and Ward. "Archaeometry" **30** (1978)

10. Private correspondence with Dr Ramsey and Dr. Hedges (Oxford) and Prof. Jull (Arizona).

11. R. Hedges in "Approfondimenti pro Sindone" **1** (1997).

12. Acts "Costa" "Argenteuil" Edition de Guibert Paris (2005).

13. B. Walsh "*The 1988 Shroud of Turin Radiocarbon Tests Reconsidered*" Proceedings of the 1999 Shroud of Turin Conference, Richmond, VA, B. Walsh Ed., Glen Allen VA: Magisterium Press (1999) pp. 326–342.

14. J.A. Christen. "Applied Statistics **43** pp 489-503 (1994).

Technical Literature:

- ✓ "Technical Repertory Mathematics & Mechanical" (Dutch Edition Elsevier) Wilcoxon test (page 1.8 f 2.8), "Kruskal-Wallis" & "Bonferoni Pairwise T-test Comparison".
- ✓ McCall "Linear Contrasts" Quality Control" July 1960.