Lessons Learned from Lord Rayleigh on the Importance of Data Analysis

Russell D. Larsen[1]
Texas Tech University, Lubbock, TX 79409

Lord Rayleigh[2], John William Strutt (1842–1919), is one of the undisputed giants in the history of physical science, having published 446 papers on a variety of topics that focused largely on wave phenomena within the field of acoustics, electricity and magnetism, hydrodynamics, optics, solids, and mathematics (*1*, *2*). Rayleigh may be best known for his explanation of the blue color of the sky, surface waves in elastic solids (Rayleigh waves), and the Rayleigh–Jeans law (a special case of the Planck radiation laws).

Rayleigh's most celebrated achievement, however, was the discovery of argon, for which he received the Nobel Prize in 1904. It is the data analysis associated with this discovery that is the subject matter of this article. In fact, I believe that the inferences drawn by Rayleigh were so stunning that they should be held as models for scientific inquiry, as they epitomize the quintessence of the scientific method.

The "lessons learned" are:

(a) accurate data are important
(b) accurate data are not enough—careful data analysis must follow data collection.

This article discusses Rayleigh's early gas density measurements, then his experiments on the density of gaseous nitrogen-containing compounds, then his data analysis, and finally, focuses on some modern (and powerful) methods of data analysis that dramatically showcase Rayleigh's inferences.

The Determination of the Densities of Gases

Rayleigh expressed an interest in carefully ascertaining the densities of gases as early as 1882 in an address to the British Association in which he said (*3*): "the time has perhaps come when a redetermination of the densities of the principal gases may be desirable—an undertaking for which I have made some preparations." Basically, Rayleigh wanted to know whether oxygen had a density exactly 16 times that of hydrogen. R. J. Strutt[3] makes a wonderful statement in his biography of his father: "Although it is difficult to argue the matter in a cogent way for those who are not in sympathy with the scientific spirit, experience gives ample proof that the labour spent in fundamental determinations of this kind does not fail of its eventual reward in scientific progress."—itself, a "lesson learned".

Rayleigh undertook his gas density measurements with extraordinary experimental care. He took into account the correction for the buoyancy of air (improperly considered by Regnault in 1845) and constructed "an inner chamber with water-proofed brick walls built for the balance, and the atmosphere in it was kept dry by the simple expedient of placing a large well-dried woolen blanket[4] in it (which) would often gain 2 lb. in weight from the moisture absorbed in twenty-four hours." Many minute experimental obstacles associated with leaks, temperature, and purification of the gases were overcome. After three years of work the first publication (*4*) by Rayleigh on the relative densities of oxygen and hydrogen appeared in 1888—his 146th publication. Other related work followed (*5*).

Rayleigh's Anomaly: The Saga of the Discovery of Argon[5]

Rayleigh next turned to the task of measuring the density of gaseous nitrogen, which he obtained from air after removal of oxygen with red hot copper and removal of hydrogen (should any exist) with copper oxide. In order to confirm the resulting density value, he also prepared nitrogen by passing air through concentrated ammonia, then over hot copper and a drying material. Ammonia gas decomposes to hydrogen, which then reacts with oxygen in the air sample, leaving additional dry nitrogen gas. A celebrated discrepancy of only 2.3 mg resulted between the two methods. Rayleigh's experimental skill was such that he was confident the discrepancy was not experimental error (which was believed to be 10 times less)—another "lesson". The "ammonia nitrogen" was definitely lighter than the "atmospheric nitrogen". Rayleigh sent a note to *Nature* (*6*) on this result "inviting criticism from chemists who might be interested in such questions". William Ramsey read the note and in a letter to Rayleigh admitted that he, too, was puzzled by and did not know the origin of the discrepancy.

Next, nitrogen was prepared by passing only pure oxygen through concentrated ammonia as before—this magnified the discrepancy to 10 mg—the "atmospheric nitrogen" was now about ½% heavier. Other methods were then tried: the reduction of both nitrous and nitric oxide each gave the same weight as "ammonia nitrogen"; so, too, did purification with hot copper as well as purification with freshly precipitated ferrous hydrate. He then tried the decomposition of urea followed by hot iron purification and finally, decomposition of ammonium nitrite, which did not require hot iron purification. After two years of work Rayleigh accepted the conclusion ("That, I take it, is a fact") that nitrogen of "chemical origin" was different from "atmospheric nitrogen".

Analysis of the Nitrogen Data 100 Years Later

The saga continues, and it is a long and fascinating saga, at that. This article is not intended to be an historical account of the discovery of argon—that the discrepancy between the two different sources of nitrogen, atmospheric and chemical, was conjectured and later proved to be due to the third most abundant constituent in dry air, argon. The full saga is best found in the aforementioned chapter on Rayleigh's son's biography (*7*).

The point to be made here is this: careful experimental work involving accurate and reproducible data collection—

[1] Present address: Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, PA 15261.

[2] The Lord Rayleigh referred to in this article is the third Baron Rayleigh.

[3] A prolific physicist himself and author of over 321 publications, Robert John Strutt became the fourth Baron Rayleigh after the death of this father.

[4] R. J. Strutt conjectured that Clerk Maxwell may have invented this method!

[5] The interested reader is encouraged to read Chapter XI of ref *3*, entitled, "The Discovery of Argon". It is a beautiful account of this scientific detective story.

Table 1. Original 15 Data Points Obtained by Rayleigh Corresponding to Weight in Grams of Nitrogen Gas from Four Sources, the "Air Source" Purified by Two Different Methods, and the Combined Chemical and Air Data Sets

	nitric oxide	nitrous oxide	ammonium nitrite	air/hot Fe	air/Fe hydrt	chem/ combined	air/ combined
1	2.30143	2.29869	2.29849	2.31017	2.31024	2.30143	2.31017
2	2.29890	2.29940	2.29889	2.30986	2.31010	2.29890	2.30986
3	2.29816	—	—	2.31010	2.31028	2.29816	2.31010
4	2.30182	—	—	2.31001	—	2.30182	2.31001
5	—	—	—	—	—	2.29869	2.31024
6	—	—	—	—	—	2.29940	2.31010
7	—	—	—	—	—	2.29849	2.31028
8	—	—	—	—	—	2.29889	—

although important—is not enough! Analysis of the experimental data is crucial. "Data analysis" implies two things: a qualitative assessment of the results and their significance (exploratory data analysis) and a quantitative mathematical–statistical treatment of the numercial data (confirmatory data analysis). I wish to emphasize the importance of the former here and to show how one can look at Rayleigh's gas density data to reach the conclusion that he reached—that the atmospheric nitrogen and chemical nitrogen samples are significantly different.

Rayleigh's own "data analysis" was actually quite primitive. Despite his exceptional experimental skill, caution, and confidence in the validity of his results, he apparently merely compared the means of the weights of his various "chemical nitrogen" samples with the means of the weights of his "atmospheric nitrogen" samples. That is all—but it was sufficient! Surprisingly, however, (and even a bit disappointing) is that there is no evidence in his work of any plotting or graphing of any kind.

Reproduced in Table 1 are the data reported by Rayleigh (*8*). The manifestation and detection of Rayleigh's anomaly is a stunning example of the efficacy of exploratory data analysis (*9*). In fact, Tukey (*10*) has shown clearly the optimal use for schematic box plots—comparison of two or more batches—the essence of Rayleigh's data. Rayleigh recognized by comparing means alone that he had two discrepant batches—the differing chemical and atmospheric samples. Table 2 presents the summary statistics for the five different samples (two for air purified by the two different methods)—a total of 15 data points, as well as for the two combined sets of nitrogen of chemical and atmospheric origin, respectively. The small standard deviations are noteworthy.

Figure 1 shows a comparison of the box plots for these two batches of data (*11*). Separate box plots are shown in Figure 2. As discussed in refs *9* and *10*, a box plot envelops the middle half of the data within the box, which contains the median value as a horizontal line within the box. The top and bottom of the box correspond to quartiles. The top and bottom whiskers go to the extreme values of the data set.

Table 2. Analysis of Rayleigh's Original 15 Data Points

X_1: nitric oxide

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.30008	.00182	.00091	3.30962E-6	.07909	4	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	1
2.29816	2.30182	.00366	9.20031	21.16144	4	

X_2: nitrous oxide

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.29905	.0005	.00036	2.52050E-7	.02184	2	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	2
2.29869	2.2994	.00071	4.59809	10.57122	6	

X_3: ammonium nitrite

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.29869	.00028	.0002	8.00000E-8	.0123	2	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	3
2.29849	2.29889	.0004	4.59738	10.56795	6	

X_4: air/hot Fe

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.31004	.00013	.00007	1.79000E-8	.00579	4	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	4
2.30986	2.31017	.00031	9.24014	21.34505	4	

X_5: air/Fe hydrt

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.31021	.00009	.00005	8.93333E-9	.00409	3	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	5
2.3101	2.31028	.00018	6.93062	16.01116	5	

X_6: chem/combined

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.29947	.00138	.00049	1.90216E-6	.05998	8	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	6
2.29816	2.30182	.00366	18.39578	42.3006	0	

X_7: air/combined

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
2.31011	.00014	.00005	2.03476E-8	.00617	7	
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**	7
2.30986	2.31028	.00042	16.17076	37.35621	1	

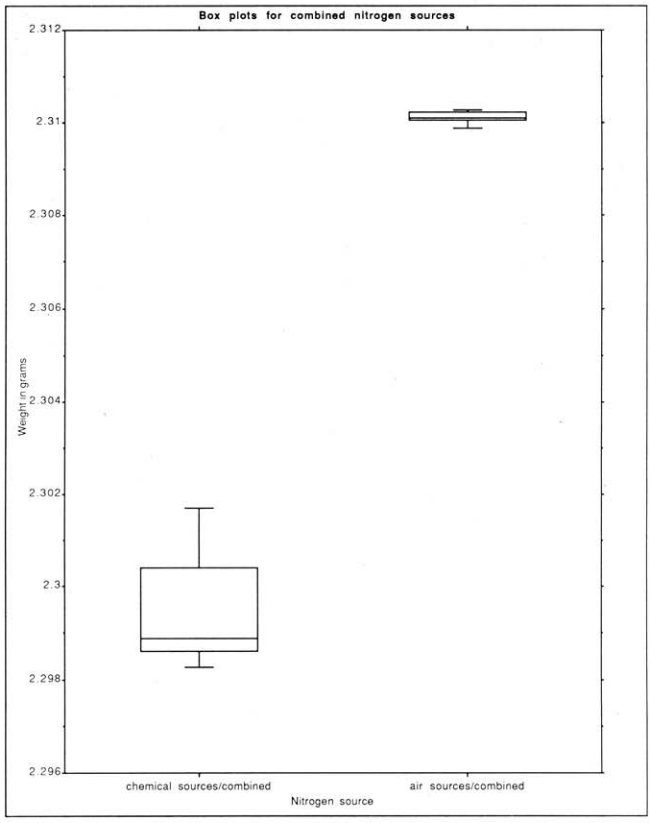

Figure 1. Comparison of box plots for combined "chemical sources" of nitrogen gas with combined "air sources" on the same scale.

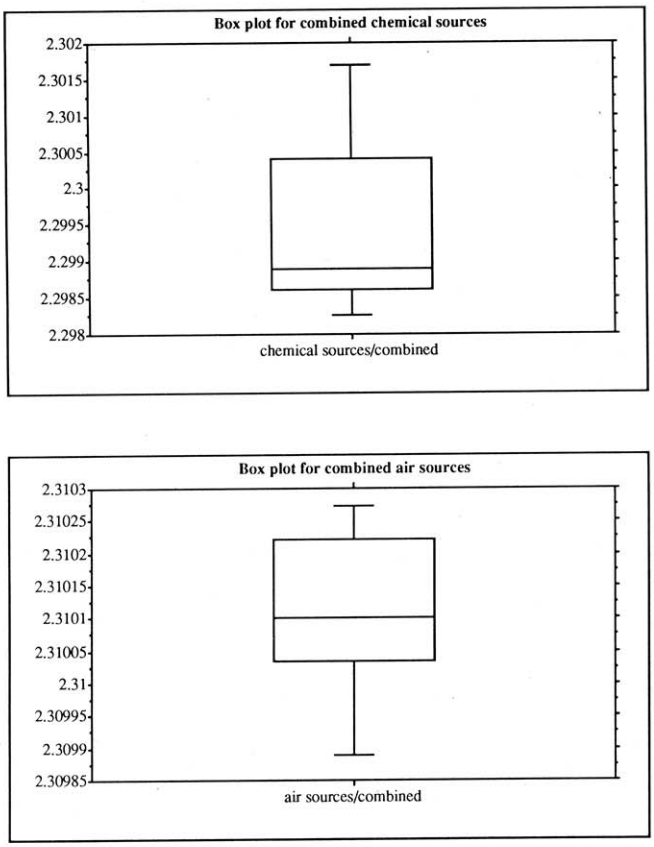

Figure 2. Comparison of separate box plots for "chemical" and "air" sources.

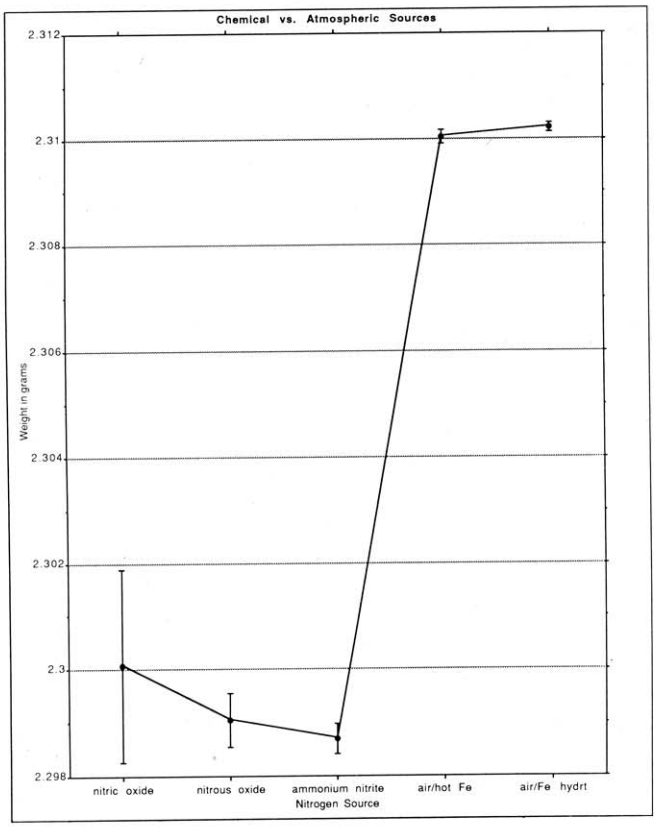

Figure 3. Graph of weight in grams of equivalent volumes of chemical and atmospheric sources of nitrogen gas obtained by Rayleigh. Air sources purified by different methods are significantly heavier.

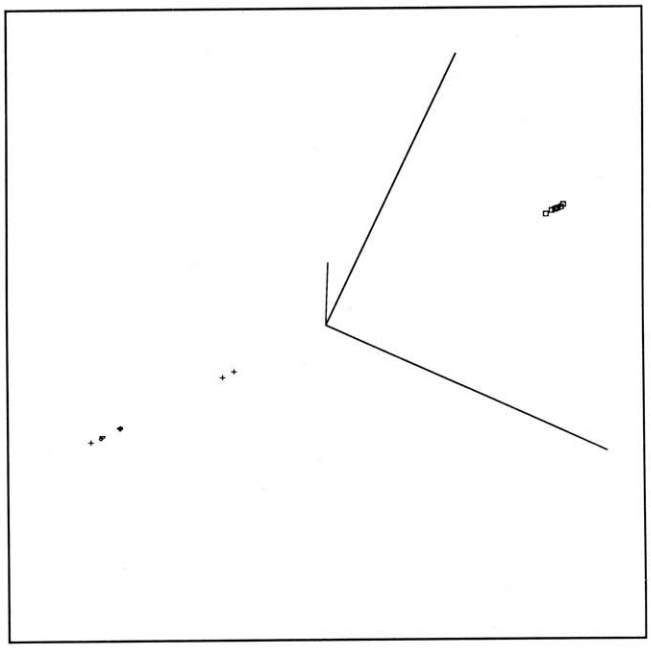

Figure 4. Three-dimensional plot of weights of 15 nitrogen-containing samples obtained by Rayleigh, plotted about centroid of data points. Note clustering at right showing the significantly heavier weight of "air set".

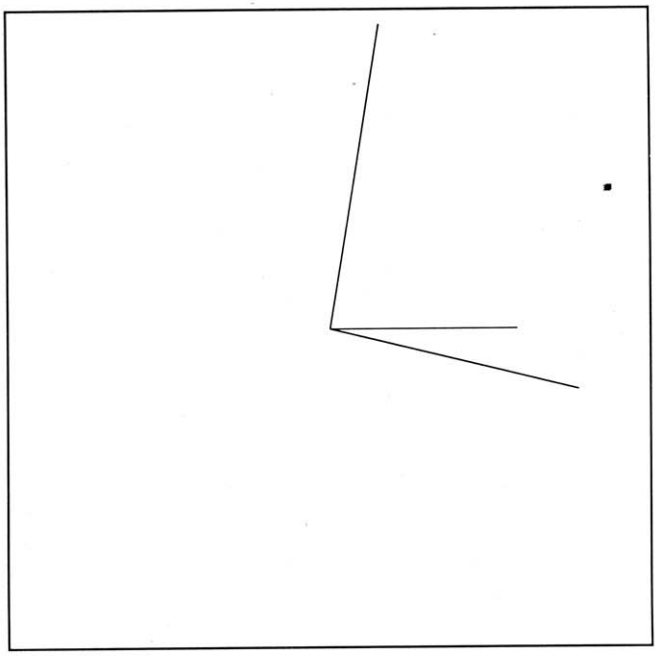

Figure 5. Data used in Figure 4 but plotted from the origin wherein obvious clustering into heavier set is not manifested.

A graph of the weights (in grams) possessed by the individual nitrogen sources is shown in Figure 3. A three-dimensional representation (*12*) of the different sources is shown in Figures 4 and 5. A plot about the centroid of the data dramatically shows the presence of two batches (Fig. 4), whereas a conventional plot from the origin (Fig. 5) does not manifest visible clustering. The appearance of the "air" cluster, and its separation from the other "chemical sources", through the use of a centroid plot (*13, 14*) is especially striking and is a relatively unknown plotting technique with obvious merit.

Discussion

The saga of Lord Rayleigh's discovery of argon provides a number of useful, if not crucial, lessons for beginning students: the importance of careful, reproducible observation, experimentation, and data collection and the commonly neglected but equally important use of critical data analysis. Without the latter, Rayleigh still would have given science his superb data sets on the densities of the nitrogen-containing gases, but these would have had no more significance than his earlier data sets on the other principal gases.

Today, we have a multiplicity of useful tools, implemented as software packages, that enhance our ability to visualize patterns and trends within data and thereby assist in interpretation of these data. With the exception of a handful of lab manuals, however, data analysis is neglected in our beginning courses and discussed far too perfunctorily for students to appreciate the importance of (and reason for) data analysis following data collection. The Rayleigh data are a superb example of the lessons to be learned. In fact, these lessons are at the crux of all laboratory sciences, not just the chemical sciences.

Literature Cited

1. Lindsay, R. B. *Lord Rayleigh—The Man and His Work*; Pergamon Press: Oxford, 1970.
2. Strutt, R. J. *Life of John William Strutt, Third Baron Rayleigh*; Univ. Wisconsin: Madison, 1968.
3. Ref 2, p 158.
4. Rayleigh, Lord. *Proc. Roy. Soc.* **1888,** *XLIII*, 356. The important result of this work was that the ratio of the densities of oxygen to hydrogen was 15.882, contradicting the supposition that oxygen had an atomic weight of 16.
5. Rayleigh, Lord. *Proc. Roy. Soc.* **1892,** *L*, 448 (publication 187).
6. Rayleigh, Lord. *Nature* **1892,** *XLVI*, 512 (publication 197).
7. Ref. 2.
8. Rayleigh, Lord. *Proc. Roy. Soc.* **1894,** *LV*, 340 (publication 210).
9. Larsen, R. D. *J. Chem. Educ.* **1985,** *62*, 302.
10. Tukey, J. W. *Exploratory Data Analysis*; Addison–Wesley: Reading, MA, 1977; pp 49–53.
11. StatView 512+ (Brainpower: Calabasas, CA) and StatView SE with Graphics (Abacus Concepts: Berkeley, CA) were used for these analyses.
12. MacSpin, Graphical Data Analysis Software (D^2 Software: Austin, TX, 1985) was used for this analysis.
13. Ref. 12.
14. Mosteller, F.; Tukey, J. W. *Data Analysis and Regression*; Addison–Wesley: Reading, MA, 1977; p 65.