The iris data set
In search of the source of *virginica*

The iris data set is one of the best-known and most widely used data sets in statistics and data science. But the origins of at least part of the data have been something of a mystery for decades. **Antony Unwin** and **Kim Kleinman** believe they have traced the source

If you have ever taught or learned statistics, data science, or machine learning, chances are good that you will have encountered the iris data set. It comprises four measurements on each of 50 plants of three different species of iris. It was first used as an example by R. A. Fisher in 1936,[1] and can now be found in multiple online archives and repositories, as well being included as part of the *base* package for the R programming language.

In its first appearance, the iris data set was used to illustrate the then new technique of linear discriminant analysis. Since then, it has been analysed using *k*-means clustering, hierarchical clustering, principal components, linear regression, logistic regression, random forests, naive Bayes classification, support vector machines, neural networks – indeed, is there any method that has not been applied to the iris data?

It is clearly a useful data set, but the source of at least part of it is somewhat shrouded in mystery. All the data were collected by the botanist Edgar Anderson, and we know that measurements for two of the three species were made in 1935, when Anderson found abundant colonies of *Iris versicolor* and *Iris setosa* at the same location. Fisher then chose to add one of Anderson's earlier *Iris virginica* data sets to these data. The *I. virginica* data were collected earlier than the others and at a quite different location. But when and where exactly?

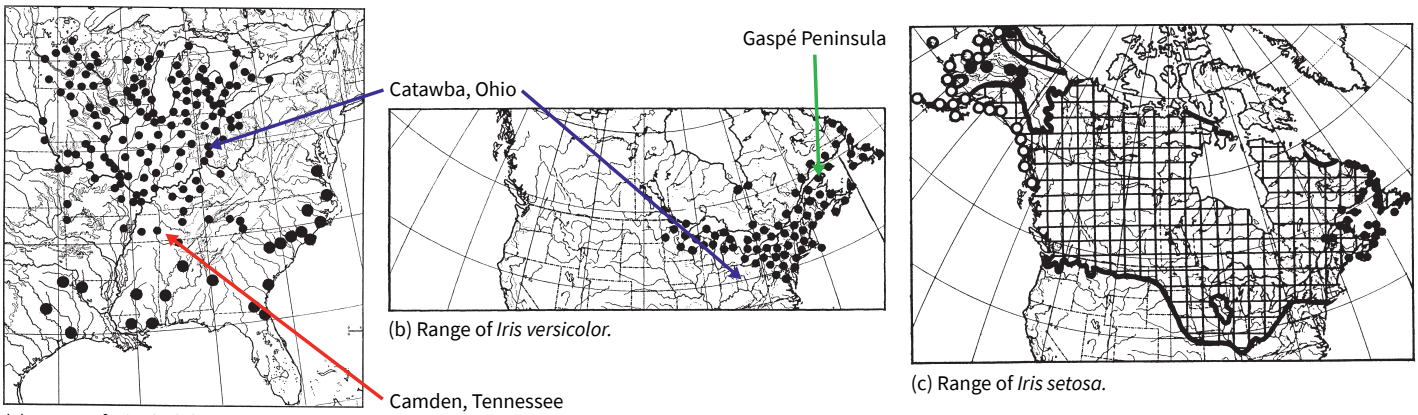
That question has long been unanswered. We believe we have found the source.

The iris species

The three species of iris that make up the iris data set are shown in Figure 1. The plants were found and measured by Anderson, who collaborated not only with Fisher but also with John Tukey. The dedication in Tukey's famous book on *Exploratory Data Analysis* reads: "Dedicated to the memory of Charlie Winsor, biometrician, and Edgar Anderson, botanist, data analysts both, from whom the author learned much that could not have been learned elsewhere."

Anderson chose to study irises because, as he wrote in 1928, "if we are to learn anything about the ultimate nature of species we must reduce the problem to the simplest terms and study a few easily recognized, well differentiated species".[2] This work opened his contributions to the breakthroughs in

Figure 1: *Iris virginica*, *versicolor*, and *setosa*. The *virginica* and *versicolor* pictures are by courtesy of the Missouri Botanical Garden and the *setosa* picture is by Tiia Monto.



(a) Range of *Iris virginica.*

(b) Range of *Iris versicolor.*

(c) Range of *Iris setosa.*

Figure 2: The ranges of (a) *Iris virginica*, (b) *versicolor*, and (c) *setosa* according to Anderson's 1936 paper.[3] In (a), large circles show *I. virginica* and small circles show *I. virginica* var. *Shrevei*. In (c), open circles are *I. setosa*, small solid circles are *I. setosa* var. *canadensis*, and large solid circles are *I. setosa* var. *interior*.

evolutionary biology over the next two decades known as the "Modern Synthesis". In this same paper he reported collecting and measuring *I. versicolor* plants between 1923 and 1928. Unexpectedly, he concluded that there were actually two species, which were initially difficult to tell apart, *I. versicolor* and *I. virginica*, the first found primarily in the north of the eastern USA and the second primarily in the south. Further study convinced him that while they were alike, they were too different to be directly linked. In a 1936 paper summarising his work on irises, he suggested *I. versicolor* could be the result of hybridisation between *I. virginica* and another species of iris, and he explained in detail why he believed an obvious candidate to be *I. setosa*.[3] He came to view such repeated backcrossing, introgressive hybridisation, as an important evolutionary mechanism.

Figure 2 reproduces Anderson's maps of the ranges of the three species.[3] *I. virginica* was to be found from "Virginia southward along the Atlantic Coast". *I. versicolor* grew "from Labrador to Winnipeg and southward to central Wisconsin, northeastern Ohio, and northern Virginia." Two varieties of *I. setosa* could be found in Alaska and one in the northwest from Labrador down to Maine. The maps are of different locations to different scales and were marked by hand. Much work was needed to collect the plants, take the measurements, draw up the tables, and make the maps.

Two of the species, *I. versicolor* and *I. setosa*, were found by Anderson growing

The plants were found and measured by Edgar Anderson, who collaborated with both R. A. Fisher and John Tukey

together at the Gaspé Peninsula in Quebec in the summer of 1935. He wrote that the data were "each from a different plant, but all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".[4] There is also a letter from Anderson to Fisher, dated 19 December 1935, in which he wrote that he had measured 50 of each of *versicolor* and *setosa* and was sending the data sheet of "four significant measurements as well as a color record". The information on the colours of the plants appears to have been lost.

As to the source of the data on *I. virginica*, Fisher writes in his 1936 paper: "The sample of the third species given in Table 1, *I. virginica*, differs from the other two samples in not being taken from the same natural colony as they were – a circumstance which might considerably disturb both the mean values and their variabilities."[1] So, Anderson was well aware of the importance of comparable origins for comparisons and describes a pretty well perfect piece of data collection, while Fisher rightly pointed out possible weaknesses of using data of different origins, and implies those data were not so perfect. But where did Fisher get the *virginica* data?

The detective work

In his 1928 paper, Anderson described how he collected and measured almost 2,300 iris plants from 96 different colonies of plants over several years across 21 US states and parts of Canada. That paper includes frequency tables of eight plant measurements, and part of Table I is shown here as Figure 3.

Apart from that table for petal and sepal lengths, Anderson included tables for the petal and sepal widths, anther length, crest, and petal and sepal tapers. Neither taper measurement was reported for the samples collected in 1928, so perhaps Anderson thought they were not as useful as he had hoped and stopped recording them. His paper describes his ideographs, a kind of glyph he proposed to represent the four measurements of petal and sepal lengths and widths. This suggests he found these four measurements the most useful. (A description of Anderson's graphical innovations can be found in a 2002 article by Kleinman.[5])

Anderson and Fisher met for the first time in 1929–30 and probably discussed these data. Fisher could have been given a copy of some or all of the raw data then. Later on, in 1937, after Anderson must have offered Fisher some further data, Fisher replied (10 February 1937): "Thanks for your letter of January 28, and for your offer of the iris data. I think, myself, that there is little purpose in your sending me a copy of these merely as a deposit in a biometrical laboratory where

workers might be tempted to do something with them." Of course, over the years, many "workers" have been tempted to do many things with the iris data set.

Everyone who has ever looked at the iris data set knows that the *setosa* plants are quite different from the other two species. Fisher may have wished to have data that demonstrated his discriminant analysis method more effectively than just for distinguishing between *versicolor* and *setosa*. He would have known that *versicolor* and *virginica* were quite similar and he probably had at least some of the data that were used for Anderson's 1928 paper. It would have made sense for him for this purpose to find data on 50 *virginica* plants that were found in a location or locations of similar conditions to the Gaspé Peninsula. In fact, he had another, more botanical aim, to test whether *I. versicolor* was a polyploid hybrid of the other two species.

There were 66 different colonies of *I. virginica* plants where Anderson had collected data. At one location, Portage des Sioux, he collected data in four different years (and these were each recorded as different colonies), but mostly he collected just for one year at each location. As it happened there was one location, Camden in Tennessee, where data on exactly 50 plants were collected in 1926. Comparing the frequency tables with the *virginica* data reported by Fisher, there is a perfect match, so these are almost certainly the data he used. Figure 4 shows two rows of histograms with bin widths matching the bins of the frequency tables in Anderson's 1928 paper for the four measurements used by Fisher. The top row in blue is the Camden data from 1926 and the bottom row in green is Fisher's *I. virginica* data given in Table 1 of his 1936 paper.

The four measurements of Camden plants are higher than in most of the other colonies Anderson studied in the 1920s. Figure 5 is a parallel coordinate plot of the measurement means for each of the 65 colonies with data on *I. virginica*. Approximate means were calculated by assuming that all values in a bin took the midpoint value. The lines for Camden have been coloured red and those for another site, Catawba, Ohio, blue. The lines for the *I. virginica* means from Fisher's 1936 article have been added in brown, confirming that the approximation made little difference.

Fisher's analyses might have looked a bit different had he, for instance, used 50 of the 57 plants Anderson measured at Catawba in 1925. A plot similar to Figure 5 but for the *I. versicolor* means showed that the Gaspé sample means were in the upper half of the distribution, with the petal length mean being higher than any of the means for the 24 *versicolor* colonies in Anderson's 1928 paper.

Mystery solved

Our search has led us to Camden, Tennessee, as the source of the *I. virginica* data in the iris data set. The details and the hard work surrounding classical data sets are often forgotten and sometimes unknown, but our investigation helped to remind us of

Our search has led us to Camden, Tennessee, as the source of the *I. virginica* data in the iris data set

IRIS VIRGINICA

Individuals of each colony classified according to size

Year	Town	State	Petal length in centimeters 1.9–2.2	2.3–2.6	2.7–3.0	3.1–3.4	3.5–3.8	3.9–4.2	4.3–4.6	4.7–5.0	5.1–5.4	5.5–5.8	5.9–6.2	6.3–6.6	6.7–7.0	7.1–7.4	Total number	Sepal length in centimeters 3.7–4.0	4.1–4.4	4.5–4.8	4.9–5.2	5.3–5.6	5.7–6.0	6.1–6.4	6.5–6.8	6.9–7.2	7.3–7.6	7.7–8.0	8.1–8.4	Total Number
1927	Kimborough	Ala.								7	5	*8**	5				25					4	6	*3*	8	3		1	
1927	Wiggins	Miss.							1		2	4	4		1		12						3		*6*	1	1	1		12
1927	Jackson	Miss.							8	*5*	5	2	1				21				2	7	*3*	3	5	1				21
1926	Arlington	Tenn.								7	*6*	2	1	1			17					6	*5*	3	2	1				17
1926	Huntingdon	Tenn.							2	1	2	2					7					1	1	2	1	2				7
1926	Camden	Tenn.							1	8	13	*15*	7	3	3		50				1	1	7	15	*11*	7	3	5		50
1926	Bonnieville	Ky.							1		8	*11*	11	4			35					1	5	8	*8*	9	3	1		35
1926	Elizabethtown	Ky.							1	1	3	3	3				11					1	1	1	6	2				11
1926	Stanton	Ky.					1		8	*8*	3	1					21			1	4	*7*	6	3						21
1926	Hayden	Ind.					2		2	*5*	4	3	1				17				2	5	*5*	6	1					19
1926	Anna	Ill.					2		6	*16*	5	8	1	1			39				2	4	8	*14*	8	2				38
1925	Vulcan	Ill.					1		3	4	*7*	6	5				26			1	2	2	7	*8*	2	3	2			27
1926	Vulcan	Ill.					1			4	*5*	5					15					1	3	*8*	2	1				15
1924	East St. Louis	Ill.							2	3	6	*8*	4				23			1		2	2	*8*	4	3	1			21
1925	Farmington	Ark.							4	2	*14*	8	9	1			38					6	7	*10*	7	6	1	1		38
1926	Pilot Knob	Mo.								2		4	1	1			8						2	2		3	1			8
1925	Wicks	Mo.					2	10	*10*	10	8	3					43			1	6	13	*8*	8	5	2				43
1926	Valley Park	Mo.						1	1	*11*	7	2					22			1		1	*10*	7	2	1			1	23
1925	P. des Sioux	Mo.							5	1							6				1	1	2	1	1					6
1926	P. des Sioux	Mo.					3	7	9	*11*	6	3	1				40			2	3	13	*10*	5	7					40
1927	P. des Sioux	Mo.							1	9	6	*8*	6	3			33						10	*8*	3	5	3	1		30
1927	P. des Sioux	Mo.								4	9	*6*	6		2		27						7	6	*10*	4	1			28
1924	Louisiana	Mo.						1	2	*12*	9	2					26					2	8	*9*	7	1				27
1928	Rich-Tex.	S. C.											1				1										1			1
1928	Eastover	S. C.							1	1	7	*6*		4	1		20						3	5	*3*	3	5		1	20

* The figures in italics are the class containing the median or mid value.

Figure 3: The first page of Anderson's Table I.[2]

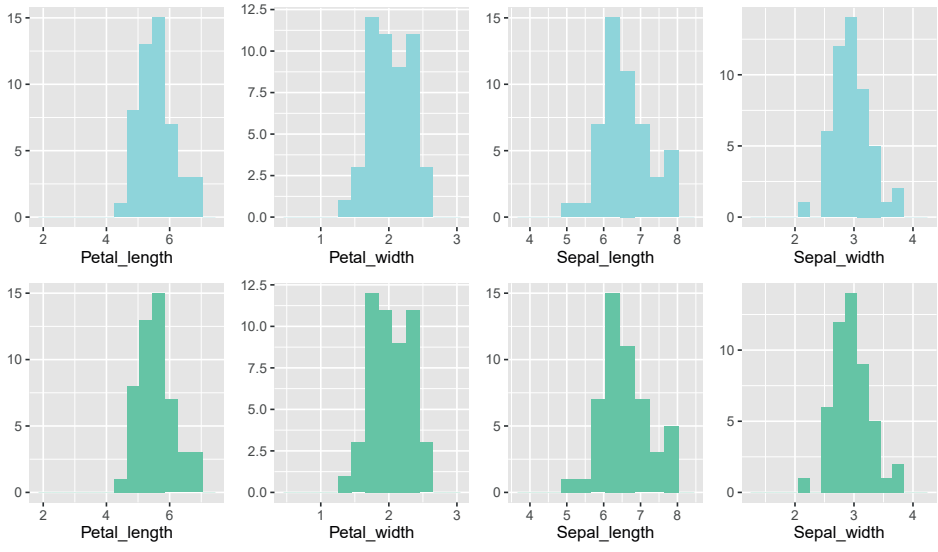


Figure 4: Histograms of Anderson's and Fisher's data for 50 *Iris virginica* plants.

Antony Unwin is a statistician and professor at Augsburg University, and the author of *Graphical Data Analysis with R* (CRC Press).

Kim Kleinman is a research associate at the Missouri Botanical Garden and is director of undergraduate advising at Webster University where he is also an adjunct professor.

Anderson's research and of the years he spent investigating irises. It also made clear that the data for the three species in the iris data set are not directly comparable because of their differing sources.

Historical data sets may be used for quite different purposes than those for which they were collected, whether they are suitable or not. Anderson and Fisher knew what they were doing with the iris data set and why. So should we. ■

Acknowledgements

Thanks to the Missouri Botanical Garden for permission to reproduce material from Anderson's paper and for the *I. versicolor* and *I. virginica* photographs, to the Fisher Archive at the University of Adelaide, and to Ulrich Fahrner and the Medienlabor at the University of Augsburg for their work in digitising Anderson's tables.

Disclosure statement

The authors declare no conflicts of interest.

Where to find the iris data set

The iris data set was one of the 100 data sets included in the *Data* book of Andrews and Herzberg that was used for many years by statisticians as a source of interesting, real data sets. Today, if you search for "iris data" on the Comprehensive R Archive Network (CRAN) you get over 1,200 hits. Not all refer to the iris data set, but the majority do. If you do the same search on the popular rseek.org website you get 264,000 hits.

The iris data set has also featured on the UCI Machine Learning Repository since 1988. In March 2021 it was the most popular data set there with just under 4 million hits since 2007, almost twice as many as the next most popular data set. Interestingly, there are actually two iris data sets there: the first is the originally submitted one that had minor errors (discussed in Bezdek *et al.*),[6] and the second, added some years later, matches Fisher's data. The iris data set has been described as the "Hello World" example of data science and machine learning.

References

1. Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

2. Anderson, E. (1928) The problem of species in the northern blue flags, Iris versicolor L. and Iris virginica L. *Annals of the Missouri Botanical Garden*, **15**(3), 241–332.

3. Anderson, E. (1936) The species problem in iris. *Annals of the Missouri Botanical Garden*, **23**(3), 457–509.

4. Anderson, E. (1935) The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

5. Kleinman, K. (2002) How graphical innovations assisted Edgar Anderson's discoveries in evolutionary biology. *Chance*, **15**(3), 17–21.

6. Bezdek, J., Keller, J., Krishnapuram, R., Kuncheva, L. and Pal, N. (1999) Will the real iris data please stand up? *Annals of the Missouri Botanical Garden*, **23**(3), 457–509.

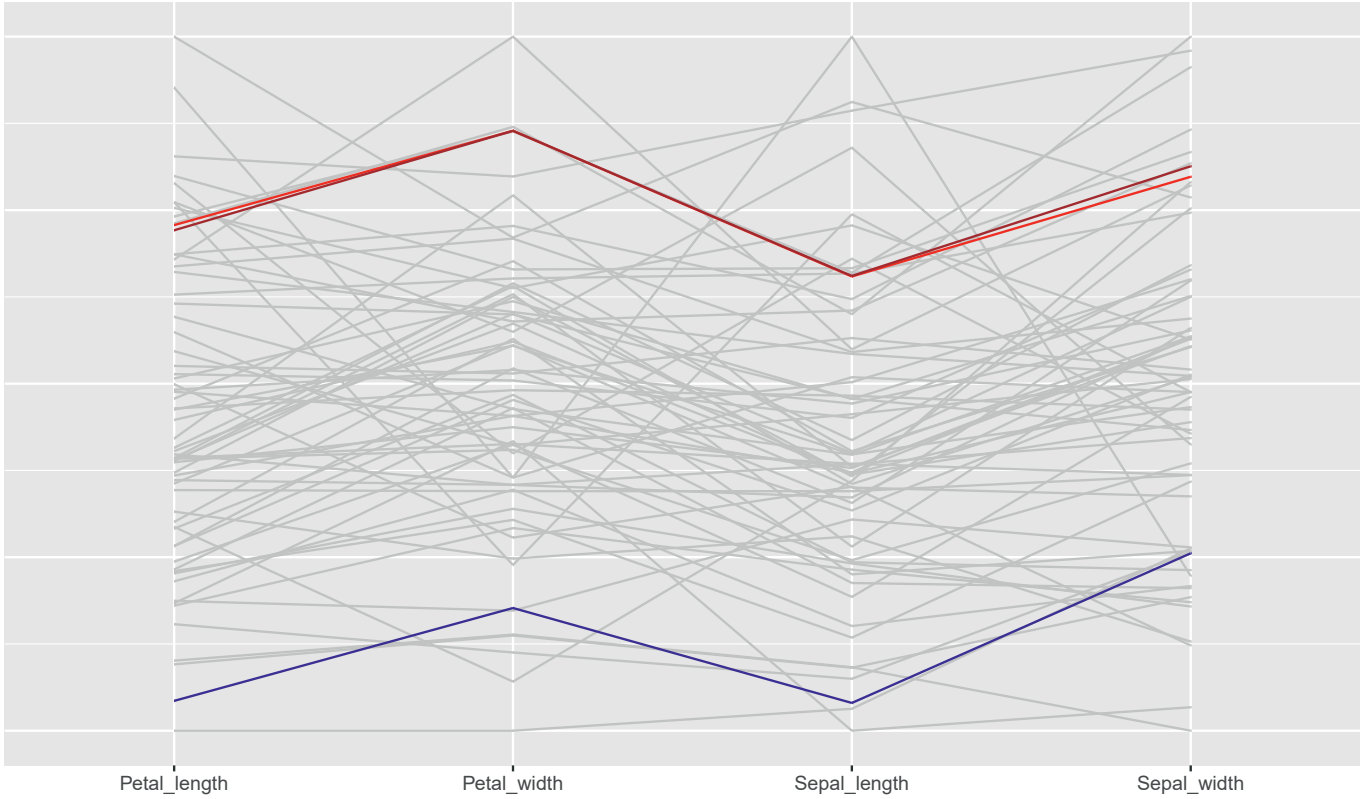


Figure 5: Parallel coordinate plot of the means of *Iris virginica* measurements at 65 colonies. The lines for Camden 1926 (red) and Catawba 1925 (blue) have been highlighted.