### 4.2.5 The Data Collection Protocol

The *data collection protocol* is the procedure for collecting and recording the data. The goal is to avoid mistakes. Michelson gives us no indication of how he planned to record his data. However, the meticulous care he showed elsewhere in the planning of his study suggests that he would have been especially careful to ensure that the data were recorded as measured.

In today's context, this step will include consideration of data entry, file structures, analysis software, and so on, especially for studies in which a large amount of data is accumulated.

## 4.3 Data

The primary purpose of the Data stage is to execute the Plan, monitoring any deviations or exceptional occurrences as they occur. Once the data are collected, processed (Michelson had to calculate the speed of light for each determination) and stored, we propose to search for anomalies and to cleanse the data set when appropriate. The goal is internal consistency. This is likely to be more profitable in an active use of PPDAC as questions about the validity of any particular value can be answered directly by the individuals making and recording the measurements.

As far as we can tell, Michelson used all of the measurements on the 1000 units that he collected. Unfortunately, he did not report all 1000 data points but instead gave the average value of the displacement $d$ for the 10 determinations in each set. All recorded explanatory variates were treated as constant over the set. The values for the measured speed of light in air for each set and the associated response and explanatory variates are given in Table 2.

[Table 2, 3 and 4 near here]

Michelson did not question the validity of his data in print. Given our computational resources, it is an easy task to examine a large number of plots of the response variates versus the order of collection or other explanatory variates not involved in the Problem. Figures xx is a plot of the recorded values for the speed of light in air versus the day of collection. Because so many values were recorded as identical, the plotted values have uniform random noise in the range from -4 to 4 added; this has the desired visual effect of spreading the points out in the plot.

8

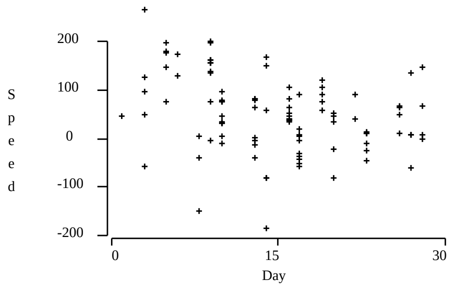


Figure 8: Adjusted speed of light (jittered) versus day.

There is an apparent decreasing relationship that is only stronger if outlying values are ignored. The noticeable exceptions to this relationship appear to be the values obtained on the last three days. However, checking with the data as presented in the table, we see that on the third last day, Michelson inverted the rotating mirror R. After two days in this position, he inverted it again to get the original position. Arguably, these changes affected the process and prior to that time the study process seemed to be drifting downwards. This also holds for morning and evening measurements considered separately. Clearly, the date is acting as a surrogate for some other lurking variable. What that could be is not known.

## 4.4 Analysis

The purpose of the analysis stage is to use the collected data and information from the Plan to deal with the questions formulated in the problem step. The form and formality of the Analysis depends on the complexity of Problem and Plan, the skill of the analyst, the amount of variability induced by the Plan, and the audience for the documentation of the study. We propose the following general breakdown of the stage:

- construct graphical and numerical summaries selected to address the Problem directly.
- model the Plan and data
- fit and assess the model
- develop formal statistical procedures

All sub-stages are directed to addressing the Problem. Michelson limited his analysis to the calculation of the average of the 100 measured velocities in air, a numerical summary and an estimate of possible error, a formal procedure. The error is based on a worse case scenario, combining probable errors based on the estimated standard deviations of replicate determinations and maximal systematic error, based on Michelson's knowledge of his apparatus and the functions used to calculate the speed of light from the measured response variates. For more discussion on the use of probable error, see Stigler [?]).

After making a small adjustment for temperature (in air) based on the effects of temperature change on the systems used to determine $\phi$, the angle of deflection, and correcting to a vacuum, Michelson concludes his analysis by reporting the speed of light in vacuo (kilometres per second ) to be

$$299944 \ \pm 51$$

Although Michelson did not formally propose a model, he carried out numerous checks that are equivalent to aspects of model assessment ([?] page 139). For example, to see if the measured speed of light was systematically influenced by the distinctness of the image, an explanatory variate, he calculated and compared the average velocities stratified by distinctness of image. This checking was repeated for many other explanatory variates.

Today, we can use corresponding graphical methods. Perhaps the speed depends on some of the explanatory variates that are not part of its calculation. For example, has the effect of temperature been successfully removed from the determinations? A plot of speed versus temperature is shown in Figure xx. A fairly weak increasing trend is discernible in the plot. However, even this trend depends heavily on the three points in the lower left corner and so is not likely to alter the result significantly. Again the values have been jittered to resolve the over-plotting of identical values.

9.

Curiously, in his comparisons of group averages, Michelson did not compare morning and evening measurements nor attempt to relate the measurement to the date. There are other curious and interesting relationships to be found in this data; we leave further exploration to the reader.

Note that there is not a clear distinction between the checks for internal consistency in the Data stage and these model checks in the Analysis stage. The same plots or summaries may appear in either stage.

Today, we can contemplate any number of ways to summarize, model and analyze the data. For example, we might construct a histogram and calculate a 5-number summary of the 100 reported values. Based on a gausssian model, which appears to fit the data well, a 95% confidence interval for the mean is

$$299852.3 \ \ 15.7$$

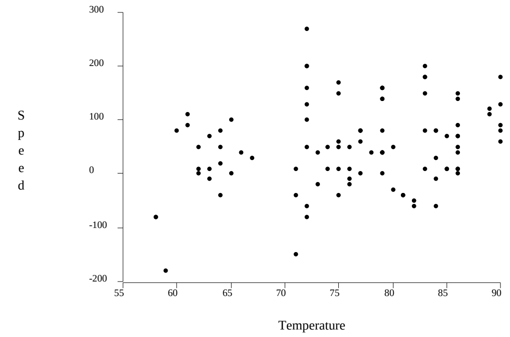


Figure 9: Adjusted speed of light (jittered) versus temperature.

Correcting for temperature, following Michelson, and converting to a vacuum, a 95% confidence interval for the speed of light (km/s) in vacuo is

$$299944.3 - 15.7$$

Note that the confidence interval is much shorter than that reported by Michelson, who included both variability and possible bias in his calculation. Other more complex modeling, analyses and model assessment can be made. The above is used to demonstrate the sub-stages within the Analysis stage of PPDAC. Again it is evidence of Michelson's precision as a scientist that his analysis so carefully parallels what can be done today.

### 4.5 Conclusion

The purpose of the Conclusion stage is to report the results of the study in the language of the Problem. As well, it provides an opportunity to discuss the strengths and weaknesses of the Plan, especially in regards to bias and variability.

In Michelson's study he concludes by reporting the speed of light (km/s) in vacuo as

$$299944 \ \pm 51$$

He then discusses possible "Objections" including among others not mentioned above, uncertainty of the laws of reflection and refraction in media in rapid rotation, retardation caused by reflection , imperfections in the lens, periodic variation in friction at the pivots of the rotating mirror and change of speed of rotation. In each case, he refers back to the Plan and the model assessment to demonstrate that the objection would have little effect on the estimate of the speed of light.

In our language, we would start with the reported speed of light based on the confidence interval. Other than the discussion given by Michelson, we would add the difference between the target and study populations as a source of possible bias.

## 5 On method in science.

When examining the writings of those who have thought long and hard about the nature of science one finds the same difficulties appearing again and again.[32] There is, for the most part, a great enthusiasm that science is progressing in some sense, that we are learning ever more about the world around us, that we are

[32] John Losee's book [?] provides a reasonable starting point.

continually solidifying that knowledge, that our increasingly sophisticated technology is testament to the power of science. Yet, when pressed, not only can we not agree on the method of science, we can't quite agree on what science is, or even whether what it talks about is real! Looking over the history described in this paper we can get some inkling as to why this state of affairs persists.

The progress seems real enough, from the question of light's speed being meaningless, to discussion of whether it is finite or not, to increasing evidence for finite speed, to ever 'better' estimates of its value. It might seem that scientific knowledge is the conjunction of the facts accumulated so far, that theories live or die according to their verification or falsification by these facts, and that, eventually, the truth will be inferred from the collection of facts.

Kuhn's work [?] describes a framework for this progress – within a scientific 'paradigm' normal science is pursued as a puzzle-solving activity, this eventually produces anomalies, anomalies accumulate until a crisis is reached, a new paradigm is somehow introduced , normal science proceeds again, and so on. For example, normal science was pursued within a paradigm where light was without speed, astronomical anomalies began to appear, leading ultimately to a theory where light had a finite speed, whereupon normal science set about solving problems to establish its value. In a more elaborate history, many such Kuhnian cycles would have been detectable.

But what about method? Long ago Aristotle wrote that knowledge, being "a state of capacity to demonstrate", required the teaching of the principles of demonstration and so the teaching of science necessarily "... proceeds sometimes through induction and sometimes by deduction"([?] $1139^b19$ - 36). But each is tricky to apply – Francis Bacon, that strongest of proponents of inductive method, allowed his perception of the incredible speed at which stars move in their orbit about the Earth to form his inductive base and so concluded that an infinite speed of light was reasonable; no lesser talents than Aristotle and Descartes by pure deduction demonstrated that light could not possibly have finite speed. Using induction and deduction in combination as in the hypothetico-deductive approach is no easier. It appears explicitly only twice in the above history – once by Aristotle to dismiss the argument of Empedocles, and once by Descartes to dismiss that of Beeckman – and wrong in both cases! At various times each of these has been suggested as *the* method of science.

A slightly different tack is to take one such method and raise it to the status of a criterion to distinguish science from non-science. Karl Popper did this in 1934 with the hypothetico-deductive approach. Contemptuous of the widely held view that the use of inductive methods distinguished science from non-science, Popper proposed instead that "it must be possible for an empirical scientific system to be refuted by experience." [33] That is, to merit the name scientific a theory must be falsifiable;[34] a decisive experiment which refutes the theory is a crucial falsifying experiment. By this criterion, the geocentric theory of the universe is scientific being falsifiable by any orbital system not centred about the Earth; Galileo's discovery of the moons of Jupiter refuted this theory. Similarly the scientific theories of light held by Aristotle and Descartes were refuted by Römer's determination of the speed of light. This criterion is turned into method by having scientists focus on trying to refute theory; theories are corroborated only by surviving the most stringent of testing.

But normal science is conservative. Crucial experiments are typically only recognized as such long after the fact – Cassini et al showed at the time that Römer's observations could be accommodated by existing theory.[35] If theories were thrown out when first refuted, the result would be chaos. Instead normal science motors along, sometimes fine tuning its theory to accommodate the new information, sometimes patching the theory with auxiliary hypotheses, and sometimes just tossing the information into the back seat where Popper's refutations become Kuhn's anomalies. As the anomalies accumulate, the ride gets rougher and some members of the scientific community become increasingly uneasy that a crisis is around the corner.

[33] [?], page 41.

[34] In a paper meant to be a general resource [?], I.J. Good gives partial prior credit to R.A. Fisher since tests of significance [?] predate Popper. This credit seems misplaced – Popper uses falsifiability as a *demarcation criterion* for science, Fisher does nothing of the sort.

[35] See [?] pages 71 - 90 for further examples and discussion.

It is here that Kuhn's work is most interesting and most troublesome. Kuhn likens the transition from one paradigm to the next to that of a gestalt shift in visual perception. Like a gestalt shift, a paradigm shift is sudden and without reason. Unlike a gestalt shift, a paradigm shift does not allow the scientist to switch between paradigms; no neutral third viewpoint exists from which both paradigms can be seen – if there were then this would be the new paradigm. This is not to say that the new paradigm cannot be reasoned about and justified to some satisfaction, but rather that it may not be possible to do so by comparing it to the old. For once the transition is complete, the convert's view of the field will have changed – its methods, its concepts, its questions, even its data – and the old paradigm can only be viewed from the perspective of the new. In a word, the two paradigms are incommensurate. Concepts, theory, methods, and data that are meaningful according to one might not be according to the other.

Consider the concept of light. According to Aristotle, light required an intervening transparent substance (like air or water); it could not exist in a vacuum. Things are transparent, of course, only because they contain a 'certain substance' which is 'also found in the eternal upper body' (possibly aether? itself a concept Aristotle tells us he has changed from that of Anaxagoras[36]). 'Of this substance, light is the activity.' But it is not movement. Moreover, the visibility in the dark of bioluminescent plants and animals does *not* depend upon light! [37] From this Aristotle says he has explained light. Not only is Aristotle's concept different from ours, but to really understand what he means by light we would need to become immersed in his paradigm. Scientific concepts like light change in irreversible ways; some like aether disappear altogether – even after thousands of years of service.

Nor are concepts alone determined by the paradigm. So too are the 'empirical facts' – Francis Bacon's data included fantastic speeds for the movement of the stars about the Earth; Glaseknapp demonstrated that different theory produced different 'observed' speeds of light. Even relatively raw 'sense data' can be dependent upon theory. Soon after Galileo announced the discovery of Jupiter's moons, he had others verify his observations using his telescopes. Many could not see the satellites; those who could see multiple lighted spots could not be certain that these were not artefacts of the new instrument. Only once the optics of telescopes was developed could there be confidence in the verity of the observations.[38] Modern instruments produce observations that are irrevocably 'theory laden.'

Paradigm shifts, incommensurability, and theory laden data have all contributed to what Ian Hacking [?] calls "a crisis in rationality" – at least for philosophers of science. Is there such a thing as scientific reasoning? Are the entities with which science deals real or are they human constructs? Does it make sense to think that there is in fact an ideal truth to which science might converge?

## 6 And what of statistics?

When statisticians look at the nature of science, they see reflected the nature of statistics.[39] Deduction becomes probability theory, induction, statistical theory (e.g. pp 6-7 of [?]); scientific method is hypothetico-deductive (e.g. [?], [?], [?]??), self-evident in statistics through formal hypothesis testing and model criticism; put it together and you have, reminiscent of Aristotle, what George Box has called "the advancement of learning" [?]. But, as the previous section has shown, science is not really like that. Neither should be our understanding of statistics.[40]

Sure statistical investigation meets with the same issues raised in the previous section but it can deal with them more easily. This is because has a considerably more focussed domain of application. For example,

[36] [?] $270^b20$-25.

[37] See [?] $418^a26$ to $419^b2$ for most of the points made here.

[38] See chapter 9 of [?].

[39] A notable exception is Pearson's *The Grammar of Science* [?].

[40] Indeed, John Tukey's long battle for the legitimacy of exploratory data analysis might have been easier if there had been greater sympathy in the statistical research community for separate contexts for discovery and for justification in science. E.g. see [?].

consider the two old chestnuts of the philosophy of science – the realist/anti-realist debate and the problem of induction.

The realist/anti-realist debate concerns whether the entities of science are real or mere theoretical constructs. The primary entities of statistical investigation are the units of the *study* population and the values of variates measured on them. The units and their collective must be determined with sufficient care for it to be possible to select any individual from the collective. Sometimes considerable effort must be put into ensuring that measurement systems return reliable values of the variates they purport to measure. Within this context, statisticians become scientific realists in Hacking's sense – if we can select them and take measurements on them, they are real [**?**]; if we cannot, then statistical investigation ceases. Whether future scientific study shows the units to be composites of other more 'fundamental' units or that the variates measured are to be interpreted differently is beside the point.

As regards induction, for statistics the problem can be neatly separated into two pieces. Ultimately, interests lies in the *target* population, as it is nearest to the broad scientific concerns of the problem. This population may be infinite, possibly uncountably so, and its definition can involve phrases like 'all units now and *in the future*.' Drawing conclusions about this population will often require arguments that are extra-statistical for they will be based on the similarities of, and differences between, the *target* population and the *study* population. Such arguments may ultimately be unable to avoid assuming Hume's 'uniformity of nature' principle ([**?**] page 89) and hence what philosophers mean by the 'problem of induction.' Such weighty problems dissipate when focus shifts to drawing conclusions about the *study* population. Such is its definition that all study populations are finite in size and random selection of units to form a sample is possible. Random selection provides the strongest grounds for inductive inference. When, for whatever reason, random selection has not been employed then either the case that it has been near enough approximated, or that the sample is itself similar in its attributes of interest to the study (or target) population must be made. The latter is much like making the case for the transfer of conclusions from the *study* to the *target* population and so can be just as difficult. In either case, the arguments will to a large extent be extra-statistical.

The critical reader might suppose that the structure we propose is designed to relegate all the difficult problems to the realm of the 'extra-statistical.' But this is not sweeping them under the rug. Just the opposite. They are exposed as potentially weak links in the chain of inference about which statistics has nothing to say.[41] The five stage structure is a template for any statistical investigation and so its applicability could be regarded as a demarcation criterion for statistics. Post-hoc, the structure allows us to identify the strengths and weaknesses in the statistical argument; in some investigations, even weak arguments may be all that are available. Ad hoc, it provides a useful strategy for finding out about populations.

# 7 Conclusions

Statistics is not about the method of science with its paradigm shifts and incommensurability; it is about investigating phenomena as they relate to populations of units. As fascinating as the questions raised in Section 5 might be, they are not our questions. That is a good thing; the empirical evidence to date suggests that they may not be resolvable.

***** Should we address somewhere the question "Does it always work?" ******** For example — Three cases EDA target = study = sample, EDA target different, study=sample, EDA target, study, sample different.

***** What about a nod towards teaching?

The five stage PPDAC process with the associated language and sub-stages provides a good framework for describing investigations such as Michelson's, especially for people learning the intricacies of Statistics. More importantly, in actively planning and executing an empirical investigation, we believe that the framework is

[41] This does not preclude further statistical studies being carried out to address some of these problems (e.g. further investigation of study error)

very valuable to ensure that important issues are at least considered. And this is the case for every statistical investigation.

Karl Pearson had it almost right. Whatever the case for science, we can say that the unity of Statistics consists alone in its method, not in its material. And it is this method that should be given the broadest dissemination.

## Acknowledgements

Thanks are due to many people for many helpful discussions. They include our colleagues Greg Bennett and Winston Cherry of the Department of Statistics and Actuarial Science, astronomers Judith Irvin of Queen's University and Dieter Brookner of Kingston who pointed out Cotter's book [?] to us, and Stephen Stigler of the University of Chicago for his helpful comments on early drafts of this paper.

All quantitative graphics were produced using the Quail statistical software environment now available on the world-wide web.