Things They Don't Teach You in Graduate School[1]

James A Hanley

McGill University

Abstract

Much of what statisticians teach and use in practice is learnt 'on the job.' I recount here some of my early statistical experiences, and the lessons we might learn from them. They are aimed at those of you starting out in the profession today, and at the teachers who train you. I stress the importance of communication.

Key Words

communication; communication; communication.

James A. Hanley (`http://www.biostat.mcgill.ca/hanley`) is Professor, Department of Epidemiology, Biostatistics and Occupational Health, Montreal, Quebec, H3A 1A2, Canada. (email: James.Hanley@McGill.CA). This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and Le Fonds Québécois de la recherche sur la nature et les technologies. The author thanks Marvin Zelen for helpful comments.

[1] An expansion on some after-dinner remarks made at the Conference of Applied Statisticians of Ireland, held in Killarney, May 17-19, 2006. The article is dedicated to two former colleagues – and superb communicators – Fred Mosteller and Steve Lagakos, who are no longer with us.

2012.07.19

0.1 Introduction

I received my formal statistical training at University College Cork (1965-1969) and at the University of Waterloo (1969-973). However, most of what I teach and use in practice was learnt 'on the job.' My first four years post-PhD were with Marvin Zelen and colleagues (Kalbfleisch, Prentice, Pocock, Lagakos, Schoenfeld, Begg, *et al.*) at SUNY/Buffalo. In 1977, Marvin moved virtually the entire 'Buffalo gang' to Harvard. There I also worked with two other academics who greatly influenced how I have practiced, Fred Mosteller in Biostatistics, and Barbara McNeil in Radiology. I joined McGill in 1980.

As I did with McGill graduate students at a career day recently, I recount some of my statistical experiences. Some of these may resonate with those of you starting out in the profession, and with the teachers who train you.

0.2 Statistical Lineage, and Statistical History

Even as an undergraduate, I was conscious of my statistical lineage: my statistics professor trained under Fisher in Rothamsted, and in his classes on ANOVA he used the same yellowed datasheets on barley yields that he had collected there. He made frequent references to the other strong personalities involved in the early 20th century developments in statistical theory. Two of us [Hanley 2004; Horgan,2005], from our class of four, have become interested in some of that history. I urge today's teachers to use JSTOR and other

electronic repositories to encourage students to trace their own statistical lineage, and to interest them in statistical history and in how statistical ideas evolved. For example, why was it that in 1908 Student called his ratio z rather than t, and used a divisor of n, rather than $n-1$, to obtain a mean squared error? What has the word regression got to do with what is taught in linear models courses, and why is it that the correlation coefficient is called r rather than (say) c?

0.3 Green

I expect that even in small departments, the first teaching experience of statistics graduate students has improved since my time. In 1968, in my Master's year, I was asked to give the compulsory statistics lectures to the medical students. They were 78 men and 2 women, all about the same age as I, and admitted to medicine more on the basis of who their fathers were than on academic merit. I had no teaching or applied experience. I tried to teach the 2-sample t-test by drawing a pair of overlapping Normal distributions, which I said were often called overlapping Bell-curves because the shapes "looked like bells." "They do not", one of the rowdy and politically incorrect male medical students told everyone, "they look like a pair of ...." Teaching statistics in compulsory courses to students who are not interested in the topic continues to be a challenge, and the task should not be left to a 'green' graduate student.

0.4 Jokes, and Other Teaching Tools

One of the teachers in our Master's year was Colm O'Muircheartaigh, now in the Harris School at the University of Chicago, and vice president in the National Opinion Research Center. He told us many jokes. One of his most memorable political ones becomes less relevant each year as fewer know who Mayor Daly was. One of his best statistical ones concerned two children who lived next door to each other in Belfast, one a Catholic boy, one a Protestant girl. You still can use this story [available on request] today to teach the scientific method, confounding, and the numbers of degrees of freedom. Despite its nowadays politically incorrect tone, I also continue to use his "Statistician and the Mathematician" story, which you can find on the Web today—possibly with the roles interchanged with professions such as engineer or psychologist—with the key words "joke", "mathematician", "half" and "practical."

0.5 First On-The-Job Lesson: Exact vs. Approximate

From my PhD training at Waterloo, I retain an enormous respect for raw data —I had to collect cigarette butts to study whether people smoked more when they switched to unmarked lower-nicotine cigarettes. And, I still go back to Cox's textbook Planning of Experiments that I studied for my comprehensive exams.

My applied work, and the fact that my supervisor was friends with

Marvin Zelen (then an adjunct professor at U. of Waterloo) brought me my first job, in 1973, in cancer clinical trials, at SUNY/Buffalo. Marvin told me "it's not what you do for your PhD that matters, but what you do afterwards" and I regularly pass on this advice to my students. Stuart Pocock helped Marvin co-ordinate the Eastern Co-operative Oncology Group biostatisticians that year, and in my first week there he asked me to write a "sample size considerations" section for my first "protocol." It was a straightforward 2-arm trial, with "response to chemotherapy" as a binary endpoint. At Waterloo I had had excellent computing facilities, but—despite having read Cox's book on my own—no formal training on such sample size matters. I spent two embarrassing weeks trying to arrive at an answer, and blamed my failure to calculate the exact critical value and statistical power on SUNY/Buffalo's poor computing facilities. That's when Stuart asked me if I had heard of the Normal approximation to the sampling distribution of the difference of two proportions.

Today, I try to ensure that my biostatistical trainees do not suffer the same embarrassment that I did. Interestingly, because they need to justify the size of their research projects, our epidemiology trainees tend to be better prepared than our biostatistical ones in this regard, even if they do not understand the statistical derivations that lead to the sample size formulae. I prefer that biostatistics students not use other people's (black box) software for such calculations, but rather work out each case from first principles.

0.6 Communication

That same first year on the job, I learned an important other lesson, this time on communication (the central theme of this note). I sent my interim report on the results of a trial (a comparison of chemotherapies for advanced colon cancer) to the oncologist who was principal investigator. He phoned me (and Pocock and Zelen) to say that my analysis "had to be wrong." I had written that the "estimated median survival" in arms A and B was 7 and 9 weeks respectively. He kept saying "look at the written protocol: it can't be!" Finally, after several futile attempts to explain my analysis, and why it was correct, and how the Kaplan Meier method deals with censored observations, etc., we realized that he was referring to the *entry* criteria. One of these stated that (in order for there to be an adequate trial of the therapy), only "those patients with an *estimated* survival of more than 12 weeks" would be accepted in the trial. I remember telling him that it wasn't my problem if the oncologists couldn't estimate very well. I gather (e.g., from John Crowley at this conference) that they still can't even today. But there is an important professional lesson from my communication difficulties back then: those of you who will work in an interdisciplinary setting should be aware that the same technical term can have different meanings for different people.

A case in point is the noun "experiment." Mathematical statistics texts tend to use it loosely to cover any data-generating process, such as tossing a coin, conducting a sample survey or using a traffic counter. Mosteller

et al. (1970, p18) use it to describe "any act that can be repeated under given conditions." In science, the word implies *intentional perturbation, for the purpose of learning.* Thus, technically, a physician who tells you "a drug-company rep left me these headache medication samples: try them out if you wish" is not performing an experiment unless (s)he also asks you to report back on how well they worked.

The word "estimated" has given grief to a McGill colleague recently when a very savvy Québec politician asked her to show him the list of the *1346* elderly patients who had been hospitalized because the deductible for their prescription drug policy was increased. If you are going to report an *estimate* from a statistical model, make sure to say "an *estimated* 1346." Better still, say "approximately 1300" (or "an excess of 1300") so people will understand that you are not able to be precise, or to say which ones are the excess cases (as in the excess deaths in the 2003 heat wave in France).

In 1975, a premier medical journal published the results [(Bernard) Fisher, 1975] of what is now a landmark cancer clinical trial of adjuvant chemotherapy for breast cancer, one that even caught the attention of Time Magazine. Later that year, Marvin Zelen gave a physical demonstration to a lay Buffalo audience of the p-value from (Ronald) Fisher's exact test applied to this clinical trial. It is the most convincing and illuminating statistical demo I have witnessed. When Marvin visited us at McGill a few years ago, I reminded him of his "marbles in a Folgers coffee tin" statistical model: Of the 67 marbles he used, 12 of them were red, to denote the 12 tumours that

would ultimately recur. He asked each person in the audience to place the 67 marbles in the tin and shake them well before drawing out at random 30 to denote what might happen to the 30 women randomized to the treatment arm (in the actual trial, the cancer recurred in 11/37 of those randomized to placebo, but only in 1/30 of those randomized to adjuvant therapy). No one in the audience of approximately 50 succeeded in having as few as 0 or 1 recurrence in the 30, thereby allowing him to say, "it goes to show that it would be difficult to achieve such a good result by chance alone."

I have used another of Marvin's ways of teaching length-biased sampling, by having students— by hand—generate the distribution of waiting times for a bus, or for discharge from hospital. I doubt if *electronic* simulations (in say R or Excel) are as effective as those using the coffee tin. Likewise, it would be difficult to improve on the 'simulation'—with 750 real samples of size $n = 4$—used by Gosset to arrive at the correct form for the distribution of s, and of the $z-$ (later to become known as $t-$) ratio.

Zelen's demo, and the analysis in the article, relied on conditional inference from the counts in a 2×2 table. In this approach, by holding both row and column totals fixed, the sample-space is reduced to a one-dimensional one. There is a somewhat more logical, and slightly more powerful, unconditional test. However, it is much more difficult to communicate via a physical demonstration, to a lay—or even a statistical—audience.

My re-enactment of the demo during Marvin's recent McGill visit was also a lesson in what *not* to do when teaching. Mosteller, who had gained

considerable experience with his 1961 televised course on NBC's "Continental Classroom"[Mosteller, 1962], warned teachers to be very careful about demonstrating probabilities 'live.' When I re-enacted the "Marvin Zelen-Folgers Coffee Tin" model, I was a short of time, but nevertheless asked a few of our students to each draw (after replacement) a random sample of 30 from the (12 red and 55 blue) marbles in the tin. It may be that I had not properly mixed them, but the first student drew out 10 red and 20 blue ones. Of course, it may have been an example of what Ronald Fisher had also warned us about: "The [small] chance will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*." In any event, I should not have carried out the partial demonstration when I knew that there was not sufficient time to allow a proper randomization each time, and sufficient repetitions to convey some sense of the frequency of the different possible outcomes of the randomization.

0.7 Errors of Type III and Beyond

I was associated with one study, which made a large statistical impact in oncology. Charles Moertel, a Mayo Clinic Oncologist, asked 16 experienced Mayo Clinic oncologists to measure 12 spheres, which he had placed on a foam base, and covered with foam to simulate tumours [he included sets of duplicate objects]. In those pre-CT-scan days, many measurements in advanced cancer were made by palpation of the tumours, and measuring the

cross-sectional area. The huge measurement errors he uncovered (Moertel and Hanley, 1976) led to the Eastern Oncology Oncology Group revising its criteria for a Partial Response from a 25% reduction in the pre-treatment cross-sectional area to a stricter one of at least a 50%. I am no longer involved with oncology trials, but I understand that the criteria have been relaxed again, to reflect the smaller measurement errors associated with modern imaging methods.

I do not have time to tell the entire story about a single data-entry error (much later, in the 1990s, in the MRI- and fully computerized digital era) that caused a small drug-company's fortunes to look promising for one day, and Multiple Sclerosis patients to be overjoyed at the prospect of a new medication for their disease. This joy was dashed a day later when, at my prompting, the "Imaging Co-coordinating Center" [McGill!] of this large multi-centre trial discovered that the *pre-* and *post*-treatment scans on which the miracle regression of the MS lesion was measured, were of the brains of two *different* patients. What led me to discover the error (I was not involved in the study, and was only consulted by the neurologist at the McGill site, after the good news was found) was my experience in the (pre-SAS, pre-PC) days in SUNY/Buffalo when every computer card had to have an ID and card number, and the pair of punched cards, the one with the pre-treatment data and the one with the post-treatment data, had to go though the card-reader correctly. Human errors continue to occur today, and more sophisticated computing technology does not mean that data-analysts can

be more complacent.

0.8 Communication, Part II

I was asked to speak about my work on bivariate survival curves at an invited session at the American Joint Statistical meetings in San Diego in 1978. I was five years post-PhD but I had to be dragged from the podium after I went way past the allotted time, but was only just over half my way through my overly-busy and overly numerous transparencies.

Steve Lagakos later told me about going with Fred Mosteller to Washington to a Food and Drug Administration hearing on Red-Dye Number 40 to present their re-analysis of the data on its effect on rats. The afternoon before, in Washington, Fred had rehearsed and rehearsed; then they went to dinner, and Steve thought they were finally done for the day. But on the way back from dinner, Fred said, "let's rehearse just once more time." They rehearsed in Fred's hotel room. Fred had had his trousers ironed/pressed and didn't want to wear them in case they would get wrinkled for the big hearing the next day. So Steve had to listen to Fred rehearse his talk in his "1940s style drawers that extended down past his knees." After I heard that, I decided that if Fred Mosteller [almost 65 at the time, with 40 years' of experience in academia, and a "television veteran"] needed to rehearse his talks, then so should I!. Fred has written a piece on giving a lecture or a talk (Mosteller, 1980), and it should be required reading for everyone under 65. Even those who are past 65 can still learn something from this dean of

American statisticians.

I learned another important lesson in communication from two other Ivy League professors. I, together with Barbara McNeil, a Harvard Professor of Radiology, sent in a paper to The Journal of Chronic Diseases in 1981 on "Maximum Attainable Discrimination." I was trying to show, from data she and I had collected on every CT scan of the head ordered at the Harvard Medical Area in 1978, that—contrary to those who thought that the results of the scans were predictable from patient signs and symptoms—scans of the head were being used only for the subtle cases. I had used a saturated model, and still could not show how neurologists could be more efficient and avoid some scans. Barbara had already tried to get me to only use a maximum of 12 (rather than 35!) slides for 10-minute talks, and didn't like my written style either . She knew I didn't always listen to her advice; so she left me submit the paper without any serious editing on her part. I still remember the review, signed by Alvan Feinstein. "I stopped reading this 28 page paper at page 6, because it was so unclear. I think that what it proposes has already been done [reference to a paper of his own, on stratification], but the paper is so badly written that, instead of re-inventing the wheel, it appears that they have re-invented the ellipse." After Feinstein's review, I took writing seriously.

0.9 Not in the Clinical Trials Textbooks

At McGill, where I am in the Faculty of Medicine, I have collaborated with researchers from pediatrics to geriatrics. In the early 1980s, a physician interested in cardiovascular disease, a cognitive psychologist and an exercise psychologist, were trying to reduce the excessive physiological reactivity (as reflected in blood pressure, etc) of Type-A individuals to stress. The collaboration stands out in my memory for the fun—and the challenges—of working across disciplines and cultures (we were at three different Montreal universities), but also for two statistical 'incidents', and for allowing me to make two psychologists aware of non-parametric methods.

After we had screened a very large number of interested employees from the railway and the telephone companies, we had 36 highly-reactive men whom we were ready to randomize into three groups, psychological counseling, administered by the cognitive psychologist herself, aerobic training, and an-aerobic training (weightlifting, a type of 'control group'), with the latter two programs designed by the exercise psychologist. Being a big believer, probably since I read Cox's book, that one "matches (blocks) first and randomizes second", I used SAS to sort the 36 eligible candidates (each one identified only by an ID number) by what my collaborators believed was the most important prognostic factor, and within that by the second most important. I then assigned those in the doubly-sorted list the letters A B C C B A etc. Then, before the trial began, I was in the enviable position of being able to make the standard Table 1 in all RCT reports, with 3 columns (A B

C) of the descriptive statistics for each of the dozen important baseline variables, even before the treatments began. The three investigators compared the columns for any imbalances (using my criteria, an "*embarrassing*" difference, rather than a "statistically significant" one) between the three groups. They noted some imbalances and asked that I reorder the list slightly to reduce them. After a few iterations, i.e., when they were happy with the balance, and presumably indifferent, I asked the cognitive psychologist which group (A B or C) she would like to have. She told me she didn't like having to choose among three letters. Maybe she was looking ahead and trying to imagine how she could, in the Methods section of the paper, describe my unorthodox, and possibly biased way of having her choose which group she got. She asked me to permute the labels and re-label them "Diamonds, Hearts and Spades." Then, she told me she liked diamonds and would take the "Diamonds" group. The second investigator chose which coded group got which one of the two remaining interventions.

The next morning, a Saturday, the entire research team met with all 36 men together to explain to them what would happen in each intervention group. Each was then given an envelope telling him which group he had been allocated to. The cognitive psychologist explained that the allocations had been prepared by a 'blind statistician' who had never met any of them before then, or even seen their names.

Ten minutes later, the cognitive psychologist called the team aside to say "team, we have a problem." The groups were about to meet with the

individual intervention leaders, and go over the plans for the first session. But one of the men told her that he had joined the study because he had a lot of stress in his work, most of it generated by his overbearing boss. Moreover, when his boss found out that his employee had qualified for this interesting trial, the boss himself signed up for the trial, and was also found to be hyper-reactive and thus eligible. If they had been both allocated to one of the exercise groups, we would not have had a problem. However, the 'blind statistician' had allocated both of them to the psychological counseling group, where the therapy consisted of group sessions, where each one had to describe the sources of their stress and, as a group, explore ways to cope with them!

I had never encountered this kind of problem in my seven years in cancer clinical trials, but the pragmatist in me immediately said "no problem, let's switch one of the two to another group." However, the cognitive psychologist insisted, "once randomized, one cannot change groups." And so, the junior man of the pair, i.e., the one who probably suffered more stress in the first place, and the one who enrolled first, deferred to his boss. The boss stayed, and the junior man left the trial. Today, I would be more proactive. As it turned out, none of the three interventions reduced physiological reactivity, so (as far as we know) the insistence on an "Intention to Treat" approach did not deprive the junior man of any benefit.

0.10 The Ten-Minute Consultation

Just as I was leaving to go on holiday that August, I got a telephone call from the same cognitive psychologist, who worked at l'Université de Montréal, "outremont" from me, to ask for my help with the sample size determination for a separate project. It concerned the drop-out rate in exercise classes, which was at that time (and probably is even now) close to 50%. She told me the intervention was "a simple psychological intervention, delivered at the well-known times when competing priorities start to matter, and motivation drops." I asked what would be a worthwhile effect, and she said, "bringing the dropout rate down to 30%." I had used the Normal approximation, learned 12 years earlier from Stuart Pocock, many times in my seven years of clinical trial work, and so I didn't even have to reach for my calculator. I told her straight off that if she wanted to have about 80% power, she would need about 95 in each arm. She said "perfect, we have about 200 subjects in all."

In December, her PhD student called me about how to analyze the data, and to explain that he was getting different advice from different statisticians. Yes, they had gone ahead and run the trial with the 200 students (they were university students), but the 200 were in eight classes of 25 each, and four of these were randomized to the intervention (a psychologist came to the class every so often and delivered the intervention to the entire class) and the other four to the "usual practice."

"Everything is different now: you have a study with an n of 8, not an

n of 200" I told him. After all, imagine if these were 8 classes in statistics, with 8 different teachers: imagine how much impact the quality of the teacher has on attendance. He told me that they had anticipated this; a particular teacher was indeed teaching two sessions, but they had given the intervention to one of these, and left the other alone. One statistician had told him that he could keep the 200 by putting in 8 dummy variables for teachers. I repeated that this wouldn't get around the problem, and quoted to him from the section on "statistics for random assignment of intact classrooms to treatments" in page 23 of the authoritative [and highly readable] book by Campbell and Stanley [1966]. There, in black and white, it states, "if the intervention is carried out on intact classrooms, then these same intact units should also be the units of the statistical analysis." We continued to argue, and so we set up a meeting with him, his supervisor and me.

At the meeting, we re-discussed the true sample size, and whether it was so small as to preclude any multivariate analysis. I reassured him that the part of his PhD where he investigated the individual personality and other characteristics that predicted which individuals would drop out, was probably unaffected and was probably not subject to the constraints of the randomization to a group intervention. He showed me the average number of sessions attended by each class:

Experimental				Control			
11.1	12.2	9.4	11.7	9.6	9.2	10.3	9.7

I suggested a 2-sample t-test, with $n = 4$ and $n = 4$, i.e., $6df$. The supervisor

objected that I wasn't allowed to use the t-test since I couldn't prove that the Gaussian assumptions underpinning the t-test applied. I saw no reason *a priori* why they would not, and proceeded to calculate the t-ratio, which came out to $t[6df] = 2.15$, "ns" by conventional criteria.

That's when I suggested a non-parametric (rank) test. The psychologist asked, "Whats that?." I explained that one did not need to be a William Gosset, or a Ronald Fisher, to work out the p-value for first principles; the calculations used to arrive at the p-value did not make strong or unverifiable assumptions. And so I proceeded to show her how we could do it right there and then. One simply ranked the 8 values:

Experimental				Control			
6	8	2	7	3	1	5	4

She was a bit surprised that I used 8 for best and 1 for worst, especially when she pointed out to me that three of her 4 experimental classes had won the "gold, silver and bronze" medals for attendance (I called these 8, 7 and 6), and that the fourth one had come in second last (what I called rank 2).

We then added up the ranks obtained by her 4 experimental classes to obtain a sum of $6 + 8 + 2 + 7 = 23$, and I explained that all remained now was to determine where in the null distribution the 23 was. In other words, how often would we get a sum as good as or higher than 23, if all we did was shuffle the 8 ranks and pick four of them at random? [This is why I like to teach this test—it is so transparent]. The psychologist liked it too, and said it struck her as being much more natural and understandable

than any other statistical test she had learned in psychology. So we began to lay out the 70 ways to pick four ranks at random from the 8. There was only 1 chance out of 70 of getting 8+7+6+5=26. The chance of a 25 was also 1/70, since to have such a sum, one would have to get 8+7+6+4. There were 2/70 chances of a 24 (8,7,6,3 or 8,6,5,4) and 3/70 of a 23 (8,7,6,2 or 8,7,5,3 or 8,6,5,4). So all in all, there was a (3+2+1+1)/70 = 7/70 = 0.1 chance of a result this or more extreme. Moreover, if the journal insisted in a two-sided test, the p-value would be 0.2 [my SAS software, which uses a Normal approximation to the Wilcoxon distribution, gives a p-value of 0.15]. Either way, the psychologist was no longer a fan of the rank test.

I said that yes, it was too bad that that rank of 2 (the class who came in second last) ruined her p-value, since if the four experimental classes had swept all top four places, the p-value would have been 2/70 even with a 2-sided test [an 8,7,6,4 finish would have yielded a 2-sided p-value of 4/70 or 0.057]. So I asked her what was it that resulted in that one class doing so badly. At that point, she was resigned to the non-significant p-value and said that one has to expect "random variation", and "the luck of the draw." That's when her graduate student interjected "But professor, don't you remember? That was the class that met at 8 o'clock in the morning." She glared at him, but I couldn't restrain myself from the "I told you so! That's why Campbell and Stanley recommend using the *class* as the unit of analysis."

Even though I was about to go on vacation when she called me that

previous August, I should not have been so hasty in providing the "canned" solution.

There was a further twist to this study. Since it was December, and classes run for just one semester, the psychologist quickly saw that all was not lost: they enrolled 8 more classes that started the next semester, in January. Now, with sample sizes of 8 and 8, their trial was more resistant to the vagaries of randomization and small sample sizes. However, we had no solution for what happened next. Half way through the January-April semester, the gymnasia staff went on strike, and the exercise classes were cancelled, leaving us with a statistical gimish. Fortunately, the graduate student had collected enough questionnaire data from the 8 Fall classes to complete a PhD on individual behavior.

0.11 Not the Usual Delta

The "delta" in sample size formulae is a good way to summarize probably the most rewarding year of my professional and personal life. My family and I spent the 1992-1993 academic year in Addis Ababa, where I taught in the MPH program within the McGill-Ethiopia program funded in part by the Canadian International Development Agency. When I returned, my department chair asked what it had been like. I replied: "you know the delta in the bottom of many sample size formulae: well, after this, for me the delta will never be the same again." The article on the sisterhood method [Hanley et al., 1996] is one small example of using appropriate (statistical)

technology, and of what we take for granted, and sometimes overly fuss about, in privileged societies.

0.12 The Bigger Picture, and Other Media

The "increase the size of your series or consult a statistician" sample size story which I related some years ago in a radiology journal (Hanley, 1989) shows the limitations of the way we still teach sample size considerations, and of the unquestioning, and even misleading, p-value approach that still often passes for "statistical inference." Fortunately, these are being replaced by interval estimation, within a frequentist, and more recently, a Bayesian, approach.

Despite this gradual shift, we continue to underemphasize to the new generation the importance of looking at the larger picture. For example, when I presented my method of approximating the (non-null) standard error for the kappa statistic, one newly minted PhD statistician asked me "but do you have a test for that?" More recently, in a whimsical article for the BMJ (Hanley et al,. 2003), we pointed out that (a) the data were less than 100% trustworthy (b) we were unable to come up with appropriate comparison groups, and (c) an important 13% of our 500 subjects had not been traced. Nevertheless, prompted by the statistical review by an eager young statistical consultant for the BMJ, the editor wrote me "I'm afraid we must insist on a p-value." The Irish in me couldn't resist responding, "Statisticians are often precise about the wrong things." Likewise, even though we had reported

that only 3 of the 435 passengers from 1912 were still alive, the knee-reaction of the internet-generation statistical reviewer was that we needed to take account of the censoring and that a t-test would not be appropriate. We pointed out that even if we gave the three remaining survivors, now well into their nineties, an additional 20 years each, it would only change the mean survival by less than 2 months, and the percentiles not at all.

Despite its whimsical nature, this BMJ report received more media attention that all of my other work put together. However, the media behaviour was very instructive. First, despite the way they fawn over you, the media are not your friends. Thus, if you wish to have your message reported accurately, take the advice of my media-savvy colleague: write it out (preferably in suitable sound bites) and keep repeating it as the answer to all the questions. Second, the reaction of one print journalist was particularly interesting. She called me to get the co-ordinates of the three survivors. I told her that we only dealt it aggregates, and directed her to the Internet site where I had extracted the raw data. But it was her next sentence that was the most telling: "it's a pity you didn't find a difference, so we could have had a story." Interestingly, the other print journalist turned the "null" result into a story, by saying that "the results turned Darwin's theory of the survival of the fittest on its head." Until then, I had not considered how much extra filtering the results scientific studies undergo before they are reported to the general public, and indeed to other scientists. By the time selected instances of our 'over-exact p-values'—often reported to several decimal places—reach

the public, they have lost most of their original meaning.

0.13 Sharing our statistical methods

Some years ago, one of our graduates sent me a very amusing and useful story about two statisticians and two epidemiologists who met on a train on the way to a conference. It relates how the statisticians were able to get by with just one train ticket between them, but how the epidemiologists got in trouble when they tried to use the same method. You can now find the story on the Internet, under various guises, if you search with the keywords "joke statisticians train conference ticket." For non-statisticians, the moral of the story may well be "never trust statisticians." But, for statisticians, it is "be careful how you explain your statistical tools to non-statisticians."

0.14 Tell Them What You Said

Three useful tips on how to give a talk are "1. Tell them what you are going to say; 2. Say it, and 3. Tell them what you said." I conclude these personal statistical anecdotes by heeding the third of these.

First, the colleagues I had when I was starting my career made a very important impression on me. For younger people considering a position—think about your potential colleagues.

Second: being a good communicator, both orally (face to face, and on the phone!) and in writing, is a key element in being a good statistical

professional. We should continue to improve these skills in ourselves and in our trainees. A wise person once said that "science begins when two people communicate." The older I get, the more this principle predominates, and the less it matters what (else) I did or did not learn in graduate school.

0.15 References

Campbell, D. T., and Stanley, J. C. (1966) Experimental and Quasi-experimental Designs for [originally Educational] Research. Chicago: Rand McNally.

Fisher B., Carbone P., Economou S.G., Frelick R., Glass A., Lerner H., Redmond C., Zelen M., Band P., Katrych D.L., Wolmark N., Fisher E.R. (1975) "1-Phenylalanine mustard (L-PAM) in the management of primary breast cancer. A report of early findings." *N Engl J Med* 292(3):117-22.

Hanley, J.A. (2004). " 'Transmuting' women into men: Galton's family data on human stature." *The American Statistician* 58(3) 237-243.

Hanley J.A., Hagen C.A., Shiferaw T. (1996) "Confidence intervals and sample-size calculations for the sisterhood method of estimating maternal mortality." *Studies in Family Planning* 27(4): 220-227.

Hanley J.A. (1989) "The place of statistical methods in radiology (and in the bigger picture)." *Investigative Radiology* 24:10-16.

Hanley J.A., Turner E., Bellera C., Teltsch D. (2003) "How long did their hearts go on? A Titanic study." *British Medical Journal* 327: 1457 - 1458 (Christmas Edition).

Horgan, J.M. (2005) "At the Cutting Edge in Longford: Francis Ysidro Edgeworth." Unpublished manuscript (personal communication).

Moertel C.G., Hanley J.A. (1976) "The effect of measuring error on the results of therapeutic trials in advanced cancer". *Cancer* 38(1):388-94.

Mosteller, F. (1962) "Continental Classroom's TV course in probability and statistics." *The American Statistician*, 16(5), 20-25.

Mosteller, F., Rourke R.E.K., Thomas G.B. (1970) Probability with statistical applications. Addison-Wesley, Reading, Mass.

Mosteller, F. (1980) "Classroom and platform performance." *The American Statistician* 34, 11-17.