DISTORTION OF STATISTICAL METHODOLOGY BY THE COURTS

by

Charles H. Proctor

Institute of Statistics Mimeograph Series No. 1986

November 1990

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

MIMEO SERIES # 1986 NOV 1990

DISORTION OF STATISTICAL MEHHODOLOGY BY THE COURTS

NAME	DATE

Distortion of Statistical Methodology by the Courts

C. H. Proctor, NCSU, November 1990

1. Background

As part of the criteria for showing racial underrepresentation on a grand jury the Supreme Court in *Castaneda v. Partida* 430 U.S. 482(1977) described the testing of a null hypothesis for a proportion. This methodology was adopted and reformulated by the United States Court of Appeals For The Fourth Circuit in their review of *Moultrie v. Martin* 690 F.2d 1078 (4th Cir. 1982). We were shown a copy of the decision in *Moultrie v. Martin* when asked by the North Carolina Attorney General's office to review such hypothesis tests for some cases in North Carolina.

The data from North Carolina were initially cast in samples of a size sometimes less than 20 cases and we were being asked to calculate "standard deviations," when what we would judge was important was the exact binomial probability. However, after reading *Moultrie* we discovered that the concept of significance probability was not made explicit there, and thus what we proposed must have seemed beside the point to the attorneys. At any rate our role as consultant was not clearly defined until after considerable discussion of the objectives of the statistical analysis and until priority was returned to the hypothesis test.

What we propose in this note is to trace the message about how to conduct a test of hypothesis for a proportion from the statistics literature through the *Castaneda* and *Moultrie* cases and to speculate on how the distortion may have arisen.

2. The Supreme Court's Guidelines

It is clear that the problem arises already in *Castaneda*, in footnote 17 where the technical details are found. References are made there to three sources of literature but the pages cited in the two statistics books deal with the binomial distribution and its approximation by the normal distribution. For example, Mosteller, Rourke and Thomas (1970) do have a section entitled "Testing of a Binomial Statistical Hypothesis," on their pages 302-314 but the citation in *Castaneda* lists only their pages "130-146, 270-291."

Footnote 17 in *Castaneda* does mention "1 in 10^{140}" and "1 in 10^{25}," but their meaning to the inference is not made palatable. In fact, these numbers are incorporated in the last sentence of their respective paragraphs as: "A detailed calculation reveals that the likelihood that such a substantial departure from the expected value would recur by chance is less than 1 in 10^{140}," and "again, a detailed calculation shows that the likelihood of drawing not more than 100 Mexican-Americans by chance is negligible, being less than 1 in 10^{25}." The impression these statements seem to convey is how esoteric is the status of a "likelihood." On the other hand the corresponding standard deviations of 29 and of 6 can easily be seen to satisfy the general and more concrete rule which they offer as: "... if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist."

The "1 in 10^{140}" and "1 in 10^{25}" numbers may appear esoteric but they can be calculated directly from the binomial probability function. This function is

$$b(s;n,p) = n!p^s(1-p)^{n-s}/s!(n-s)! \qquad (2.1)$$

which gives the probability of exactly s events in n trials when p is the chance of the event happening at any one trial. I am using the notation of Feller (1957) who also provides (p. 140) an upper bound on the chance of s or less as:

$$\sum_{p=0}^{s} b(p;n,p) < b(s;n,p)\frac{(n-s+1)p}{(n+1)p - s} . \qquad (2.2)$$

For both of the above examples p = .791. The "1 in 10^{140}" results from n = 870 with s = 339 and I find the bound to be 10^{-144} and so I would say "1 in 10^{144}." The second case has n = 220 with s = 100 and yields $10^{-2\cancel{0}7}$ by my calculation.

Because the probabilities are so extremely small there is ample opportunity for rounding error but for both of these examples there is no problem in recognizing that the evidence is strong that there is a discrepancy between the observed and the expected proportions. On the other hand, because of the difficulties of calculation and the difficulty of comprehending a chance of, say, 1 in 10^{140} one can easily understand why the Court emphasized the number of standard deviations rather than the significance probability.

We might also note here that p was set to .791 for both examples because the smaller sample is in fact part of the larger. Such a practice of furnishing two conclusions or inferences from overlapping data is generally not advisable since it may create an impression of agreement over separate realities that is, in fact, not warranted.

Although the strength of evidence comes from the smallness of the significance probability, the Court manageo to shift attention to standard deviations. Let us speculate as to why this may have happened. There are certain shortcomings of empirical significance probabilities. Certainly the statistics profession has itself clouded the picture by complaining over the "rigidity" of the 5% and 1% levels and bickering over two-tailed versus one-tailed significance probabilities. More recently Bayesian statisticians have offered snide comments about involving "in the evaluation, observations that have not occurred." J. O. Berger (1985) p. 29. But these shortcomings should not detract from the basic understanding one has of a "1 in 20" or "1 in 100" event nor from the application of that understanding to any data under review.

The criticism that inference has been slavishly tied to just two levels of significance, the 5% and the 1%, was perhaps to some extent justified before computers and calculators had been developed to overcome dependence on tabled values. But now it is routine to report an empirical level of significance with several digits of accuracy. One will now report "significant at the 5.6% level," whereas earlier one may have said "not significant at the 5% level." This is also true for testing a binomial proportion, where a computer program can quickly generate all needed probabilities, and, although there may be some reluctance for the Court to enter into a discussion of computer programming, it would seem a necessary step to be taken.

There may be a further reason relating to computational ease for casting the criterion in loose and fairly conservative terms of "two or three." This involves the continuity correction. When approximating to the binomial chance of getting 100 or less in 220 trials when 174 were expected one would calculate the standard normal deviate as:

$$Z = (100 + .5 - 174) \div \sqrt{174 \times 46/220} = -12.2$$

rather than as:

$$Z = (100 - 174) \div \sqrt{174 \times 46/220} = -12.3 .$$

Omitting the continuity correction tends to exaggerate the number of standard deviations and, if one suspects that the calculation will often be done without the correction, then it may be wise to quote only the loose and conservative critical points of "two or three."

Another issue is two-tailed versus one-tailed significance probabilities. In the present case where evidence of discrimination will be just on one side, the one-tailed probability is called for. This is convenient since calculating two-tailed significance probabilities for a non-symmetric distribution can pose problems of uniqueness. In this regard, shifting to standard deviations seems to obviate this difficulty since the normal distribution is symmetric. Perhaps this consideration figured in causing the Court to emphasize standard deviations. I am here attempting to read the background and possibly unconscious considerations used by the statistical consultants to the Court, and I hope the reader will treat the reasoning as tentative.

The inclusion in the significance probabilities of values more extreme than the observed is clearly needed when the observed value is a continuous quantity and is equally logical when it is discrete. The smallness of the significance probability is used to provide strength of evidence and thus adding to this probability will weaken the evidence. If, even after adding the probabilities of the more extreme departures, the significance probability is still impressively small then the evidence is strongly against the null hypothesis. That is, adding additional probabilities assures that we will not artificially exaggerate the evidence.

The criticism that we have just refuted points to the illogicality of including in the event underlying the significance probability outcomes that did not occur. This criticism may, however, be cogent when leveled generally against the use of two-tailed significance probabilities. That is, outcomes underlying a significance probability should, it seems to me, be adjacent in a physical sense. This preference may be idiosyncratic and at present I can cite no general principles to support it.

A fourth possible reason for not preferring the significance probability may be the recognition that setting the reference proportion is not an exact operation. One may prefer to be rather loose about declaring significance, saying "two or three" rather than, for example, "2.326" which is the 1% point, in recognition of the crude value being used as the null hypothesis proportion. This problem with the reference value may, however, not exist in some applications and where it does exist it may be reasonable to stipulate some upper or lower bound. At any rate, such a source of inexactitude should be dealt with separately and not be permitted to excuse inaccuracies in other steps of the inference.

We have thus listed a number of reasons for avoiding or downplaying the usefulness of a significance probability. I believe these reasons are not compelling and that we have attempted to deal with each of them briefly. Thus, in summary, we may suspect why the Court chose to emphasize standard deviations of departure, but we hope we have shown that this was not wise. Now we will, so to speak, verify our finding by examining how the United States Court of Appeals For The Fourth Circuit has added to the problem.

3. The Fourth Circuit's Interpretation

The ***Moultrie v. Martin*** decision with which we were provided to serve as our guide in examining the North Carolina data, contains the statement "... the courts of this circuit must apply a standard deviation analysis such as that approved by the Supreme Court ... ." In addition to propagating this fairly subtle methodological error it makes a rather bold one of its own by stating: "While the Supreme Court does not explain the adjustments necessary for small sample sizes, we have discussed them in the margin[10]." Footnote 10 itself begins as: "The Supreme Court's rule in *Castaneda* of course can be adjusted for small sample sizes through the use of the student's t distribution."

The entries in Table 1 are designed to help in understanding the effects of correcting for continuity and of using the student's t distribution. Suppose there are n = 30 trials and the chance of an event is p = .4 at each trial, then 12 events would be expected. The exact chances of three, four, five, etc. are given in Table 1 along with the chances of three or less, of four or less, etc. It is the latter that one takes as significance probabilities. Notice that the uncorrected normal approximation actually does better than the normal approximation with continuity correction, when the number of cases observed is very small, but as the cumulative probability rises to 1% or 5% then it is better to be using the correction. The Student's t distribution is simply not appropriate because it has relatively thick tails.

Table 1. Comparison of Approximations to Exact Binomial Probabilities for n = 30 and Expected Number of Cases = 12.

No. of Cases Observed	Binomial Probabilities		Normal		
	Individual	Cumulative	Corrected	Uncorrected	Student's t
3	.0003	.0003	.0008	.0004	.0018
4	.0012	.0015	.0026	.0014	.0046
5	.0041	.0057	.0077	.0045	.0109
6	.0115	.0172	.0202	.0127	.0248
7	.0263	.0435	.0468	.0312	.0521
8	.0505	.0940	.0961	.0680	.1012
9	.0823	.1763	.1757	.1318	.1796
10	.1152	.2915	.2881	.2280	.2902

Table 2. Binomial Probabilities and Normal Approximation for the Case of n = 18 and p = .38.

No. of Cases	Binomial Probabilities		Normal	
	Individual	Cumulative	Corrected	Uncorrected
0	.0002	.0002	.0010	.0034
1	.0020	.0022	.0048	.0095
2	.0105	.0127	.0175	.0251
3	.0344	.0472	.0524	.0616
4	.0791	.1263	.1279	.1358

The Fourth Circuit then presented data and subjected it to the criteria suggested by *Castaneda*. For example, in 1977 there were 3 blacks among 18 jurors, while the voting rolls were found to contain 38% blacks and thus $p = .38$ became the null hypothesized value. This resulted in a calculation of -1.86 standard deviations, which the Court chose to report as 1.8. It is of interest to notice in Table 2 how close to the 5% level of significance is this result. In fact, under the null hypothesis, 3 or less black jurors has a .0472 chance of happening. The uncorrected normal approximation (using -1.86) gives a 6.16% chance and even with correction for continuity it goes only to 5.24%. The upper bound formula from Feller, see (2.2) above, gives .0496, which is also under 5%.

The full data on numbers of black jurors for the seven years 1971 through 1977 were: 1, 5, 5, 7, 7, 4 and 3. There were 7 x 18 = 126 total cases and 38% would be expected to be black, i.e. 47.88, while 32 were found to be black. However, the Court states that "twelve people are randomly chosen each year to serve as grand jurors. Six of these twelve were chosen at random to serve for a second year." This means that there were not 126 trials or choices behind these data. Actually there were 90. Six were made in 1970 and 7 x 12 = 84 were then made in the next seven years. Certain of those trials (54) only counted once but others (36) counted twice.

The variance in the observed number of black jurors is thus found to be $(54 + 4 \times 36).38 \times .62 = 46.6488$ and the discrepancy in standard deviations between observed and expected is $-2.325 = (32 - 47.88)/\sqrt{59.3712}$ which becomes -2.252 after correction for continuity. The level of significance is .012 which means that if the draw is random then the observed result would occur in less than 1 in 80 times.

Notice in this case that the distribution of the number of black jurors is not binomial so the normal approximation becomes the most practical approach. Roughly speaking the 126 trials do not satisfy the condition of independence from one to the next. This situation may be unique to South Carolina, where the *Moultrie* case arose, but it does point out the pitfalls awaiting anyone in applying suggested methods outside of routine situations.

The Fourth Circuit used a total of 31 cases observed rather than 32 and we do not know which number is correct. The Court also recalculated standard deviations on data after omitting the first year. Since all comparisons were to student's t values the Court found no grounds for rejecting the null hypothesis.

We would suggest that a single period of years and thus a single test be made and in accord with a recommendation of *Castaneda* the longest period is to be preferred. Thus we would use the significance level of .012 and note that this is fairly strong evidence for rejecting randomness of jury selection.

4. Discussion

Having seen some of the problems created by the use, in a legal setting, of such a seemingly simple method as a test of a binomial proportion, we attempted to do the tests in North Carolina so as to avoid these mistakes. Data were pooled over relatively long stretches of time to build up sample size and exact binomial probabilities were used to express the strength of evidence. It remains to be seen whether new difficulties may arise. Judging from what we have seen they should be expected.

References

Berger, James O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer-Verlag, NY.

Feller, William (1957) *An Introduction to Probability Theory and Its Applications*, vol. 1, 2nd edition. Wiley, NY.