Today’s Wall Street Journal carries an article questioning Boston Scientific Corp.’s claim of success in the clinical trial for its Taxus Liberte heart stent, which the company says is the top-selling model outside the United States. Here is the mathematical background behind the article. More background information is in Dr. Mark Turco’s 2006 presentation (http://www.europcronline.com/webcasts/2006/taxus_atlas) and Boston Scientific’s 2007 medical-journal article on the study results (Turco et al., Polymer-Based, Paclitaxel-Eluting TAXUS Liberté Stent in De Novo Lesions, J. Am. Coll. Cardiol. 2007;49:1676–83).

Q: What was the primary goal of the TAXUS ATLAS trial?

A: The primary end-point was to show non-inferiority of target-vessel revascularization after nine months, comparing Taxus Liberte against an “entry-criteria-matched” population of patients who had Taxus Express implanted in earlier trials. The primary end-point p-value was calculated by comparing the per-protocol populations (people who actually got the prescribed stent), not the intention-to-treat populations, since violations of the protocol would tend to bias the result in favor of non-inferiority.

“Entry-criteria-matched” meant that the Express patients’ records were screened to include only those who fit within the same outer limits (e.g., on vessel diameter and lesion length) as were applied to patients receiving the Liberte. There was no case-matching, multivariate propensity-score adjustment, or blinding of the Liberte patients. The Express patients had been blinded as to whether they were receiving a drug-eluting Taxus Express or a bare-metal Express stent.

“Non-inferiority” meant demonstrating that a one-sided 95% confidence interval on the difference of proportions excluded a 3-percentage-point increase in the rate of 9-month TVR. That is, the Liberte was to be considered “non-inferior” if the non-inferiority hypothesis could be rejected at the $p < 0.05$ level. (The hypothesis being that Liberte’s true rate of TVR is at least 3 percentage points greater than Express’s.)

Q: How did Boston Scientific calculate the p-value for the primary end-point?

A: The company’s statisticians calculated a one-sided Wald interval with the SAS software, then consulted a statistical table to find the corresponding p-value. The hypothesis test for whether the primary end-point was “met” can be called a z-test without continuity correction, estimating standard error from point estimates.

Each arm of the study is modeled as a series of draws from a binomial distribution with unknown $p$. E.g., the 855 Liberte patients who made up the per-protocol population for purposes of the primary end-point constitute 855 samples of a binomial random variable with unknown, but constant, $p$ — the Liberte’s true rate of nine-month TVR. The same for the 956 Express patients.

The study recorded 68 Liberte TVRs (for an observed rate of 7.95%), and 67 Express TVRs (7.01%). Adjudication of the end-points was not blinded as to which stent had been implanted.

To calculate the standard error, the Wald interval approximates each binomial with a normal distribution whose mean and variance is estimated from the observed rate of the corresponding binomial. For example, to approximate a Liberte arm of $m$ patients with $i$ TVRs, the test uses a normal distribution with mean $i/m$ and variance $i(m - i)/m^3$. The same for an Express arm of $n$ patients with $j$ TVRs.

The p-value for non-inferiority was the probability that samples drawn from two normal distributions with such parameters would differ by more than the non-inferiority margin, in one direction:

$$p = \int_{0.03}^{\infty} N \left( \frac{i}{m} - \frac{j}{n} \cdot \frac{i(m - i)}{m^3} + \frac{j(n - j)}{n^3} \right)$$

where $N(\mu, \sigma^2)$ is the probability density function of a normal distribution with mean $\mu$ and variance $\sigma^2$.

Written out explicitly, this becomes:

$$p = \int_{0.03}^{\infty} e^{-\frac{m(n)(i(m - i) + n(j - n))^2}{2m^2n^2(jm - in + nm - in^2)}} \frac{d}{dx} \left[ \frac{(i(m - i) - j(n - j))^2}{m^3 + n^3 + 2(n - j)n^2 / \sqrt{2\pi}} \right]$$

Filling in $i = 68$, $m = 855$, $j = 67$, and $n = 956$, we get $p = 0.04873952...$, just under the 0.05 cutoff.
Q: Why doesn’t this mean the results were statistically significant?

A: The problem is that in the regime of the ATLAS trial, the Wald interval consistently overstates the certainty of statistical results. With a cutoff of $p < 0.05$ for significance, the method is too easy — it allows a false-positive rate of more than 5%.

A $p$-value represents our maximum tolerance for type-I errors (false positives). A level-$\alpha$ test gives us a value of $p$, and if we reject the null hypothesis only when $p < \alpha$, we are guaranteed that our rate of false positives won’t exceed $\alpha$. That’s why receiving a $p$-value less than 0.05 from a pre-specified level-alpha test allows us to claim statistical significance: such a procedure, if repeated, will produce false-positives at a rate less than $\alpha$.

But the traditional $z$-test, used in this non-inferiority context, is not a level-$\alpha$ test. E.g., with a cutoff of 0.05 (as used in the ATLAS trial), its rate of false positives can exceed 0.05.

Here is the type-I error rate of the ATLAS test with the group sizes in the ATLAS per-protocol population, charted along the border of the non-inferiority hypothesis — the line on which the Liberte 9-month TVR rate, $\pi_{TVR_\ell}$, is 3 percentage points higher than the Express TVR rate $\pi_{TVR_e}$.

![False Positive Rate of ATLAS non-inferiority test along critical line](image)

The interval cannot be called a 95% confidence interval for any region of this critical line — its best case, worst case, and average case performance all undercover, leading to excessive false positives at every value of the parameters. Since using a cutoff of $p < 0.05$ leads to a false positive rate greater than 0.05 at every point on the critical line, this test can fairly be said to consistently exaggerate the certainty of the results. (We constrain our observation to plausible 9-month TVR rates for drug-eluting stents, which are generally less than 10%. At rates above 40%, the Wald interval no longer consistently undercovers on the critical line.)

The excessive false positives and undercoverage are partly a result of the approximation of binomial distributions by normal distributions, the lack of continuity correction, and the decision to approximate each binomial distribution separately according to the point estimate of its rate.

Such approximations do not greatly alter the $p$-value, at least when the rates are more than 5%. But close to the border of significance, passing the test can’t justify rejecting the null hypothesis at the 0.05 level, because the test’s rate of false positives isn’t bounded by 0.05.

A better approximation would be to estimate the standard error from the maximum-likelihood pair of rates that is consistent with the null hypothesis (in this case, the hypothesis that the rates differ by 3 percentage points or more). After all, the goal is to calculate the probability of obtaining the observed result, or one more extreme, if the null hypothesis were true. Adopting a value of the standard error that contradicts the null hypothesis leads us to reject the hypothesis with too much confidence.

Using the constrained maximum-likelihood estimates would still be a $z$-test, but would no longer use a Wald interval. It could be called a “$z$-test with constrained maximum likelihood” or a “score test” or, in some contexts, a “chi-square.”
Here is the same graph, but for the score test: a $z$-test with standard error estimated from the constrained maximum likelihood estimate.

The score test isn’t perfect, since its false-positive rate still exceeds $\alpha$ at some points along the border of the null hypothesis. But its performance is much closer to the desired, and where the approximation gets worse, it errs on the side of conservatism instead of lenity. The ATLAS results would have failed this test, with a $p$-value of 0.05151.

The above graphs of false-positive rates were calculated by totalling up the probability of success of the test, given $\pi_{TVR\ell}$ as Liberte’s TVR rate and $\pi_{TVRe}$ as Express’s. We assume the study design fixed the number of patients to be $m = 855$ in the Liberte per-protocol arm, and $n = 956$ in the Express per-protocol arm. There were $(m+1)(n+1) = 856 \times 957 = 819,192$ possible outcomes of the trial. The probability of a particular outcome $(i, j)$ is a product of binomial distributions:

$$P_{\ell e}(i, j) = \binom{m}{i} (\pi_{TVR\ell})^i (1 - \pi_{TVR\ell})^{m-i} \binom{n}{j} (\pi_{TVRe})^j (1 - \pi_{TVRe})^{n-j}$$

When $\pi_{TVR\ell} - \pi_{TVRe} \geq 0.03$, any outcome where the test succeeds is a false positive (a type-I error). The total probability of all successful outcomes in such a situation is the rate of type-I error. This rate is largest along the border of the null hypothesis, the line $\pi_{TVR\ell} - \pi_{TVRe} = 0.03$.

To be clear, the flaws in the Wald interval and the borderline result don’t mean the Liberte is inferior to Express — just that the Liberte’s trial failed to rule out that possibility at the conventional 0.05 level. It’s the study that failed (to reach statistical significance in the primary end-point), not the stent. Since the study was only 86% powered (that even if the two stents were identical with a 5.7% TVR rate, the study would have a 14% chance of failure), this is not that surprising an outcome.

Q: Did the medical journal article discuss the statistics behind the primary end-point?

A: No. The manuscript said, “Student $t$ test was used to compare independent continuous variables, while chi-square or Fisher exact test was used to compare proportions.” This statement appears to have referred to secondary end-points and other comparisons. The primary non-inferiority analysis used a Wald interval or $z$-test, not a chi-square or Fisher exact test.
Q: How close did the results come to failing the Wald test?

A: As close as possible. A single extra TVR in the Liberte group, or one fewer TVR in the historical Express group, would have led Boston Scientific to have to announce a failure to meet the primary end-point. Here is a chart showing the \( p \)-value and significance result for outcomes in the vicinity of the actual outcome (circled), \((i,j) = (68, 67)\):

\[
\begin{array}{cccccc}
70 & 9.7 & 8.4 & 7.2 & 6.2 & 5.3 \\
69 & 8.1 & 7.0 & 6.0 & 5.1 & 4.3 \\
68 & 6.7 & 5.7 & 4.9 & 4.1 & 3.5 \\
67 & 5.5 & 4.7 & 3.9 & 3.3 & 2.8 \\
66 & 4.5 & 3.8 & 3.1 & 2.6 & 2.2 \\
\end{array}
\]

TVR (Liberte)  TVR (Express)

The fact that the study landed right on the edge is part of why the use of the Wald interval led to a false sense of success. If the study had landed farther into successful (or failure) territory, the boost provided by the Wald interval’s undercoverage wouldn’t have been enough to bump the study over the line from failure into success. It’s only because the study was already the ultimate borderline finding that such sensitive statistical issues come into play.

Q: Did Boston Scientific follow the pre-specified Statistical Analysis Plan?

A: There’s no evidence they diverged from it, and Boston Scientific says it followed the plan exactly. The company and the FDA have declined to release this document, so it’s not possible to verify that assertion independently. The company has released an excerpt from the plan that calls for a “Two-group Z-test of non-inferiority in proportions (normal approximation to binomial)”.

The excerpt doesn’t clarify exactly which \( z \)-test was to be used — a Wald interval method (which the company ended up using) or a \( z \)-test with standard error estimated under the assumption that the null hypothesis is true. There is no evidence to cast doubt on the company’s assertion that it pre-specified the use of the traditional standard-error estimate elsewhere in the Statistical Analysis Plan.

Q: If the test was pre-specified, doesn’t that immunize the results against criticism that a different test should have used?

A: No, pre-specification is necessary, but not sufficient. The reason that a hypothesis test must be pre-specified is that the goal is to construct an experiment whose long-run false-positive rate is less than some specified \( \alpha \), in this case 5%.

Each hypothesis test designates different “successful” outcomes and “failure” outcomes, and is constructed so that if the null hypothesis is true, the probability of landing on a “successful” outcome (a false positive) is less than \( \alpha \).

If the hypothesis test could be chosen after-the-fact among all available options, we would instead have access to the union of successful outcomes among all the options. Any outcome that was designed a success by the maximum-likelihood standard error \( z \)-test, the \( z \)-test with Yates continuity correction, the Agresti-Caffo \( I_4 \) test, the Farrington & Manning score test, the Miettinen & Nurminen score test, or the
double-binomial exact test with any Berger-Boos \( \gamma \) parameter for restricted range would become a successful outcome for us.

Even though the available tests each individually satisfy the constraint on false-positive rates, if we let the union of their successful outcomes all be successful, then we would no longer satisfy the constraint. That’s why pre-specifying the hypothesis test is necessary.

But it’s not sufficient: if the test we pick doesn’t keep false positives at less than a 5% rate, we can’t say we’ve rejected the null hypothesis at the 0.05 level just by passing the test, or that a 95% confidence interval excludes the null hypothesis. If following the procedure of claiming a success only when the equation’s output is below 0.05 doesn’t bound the rate of type-I errors by 0.05, then we can’t formally call the equation’s output a “\( p \)-value.” In the ATLAS trial, 0.04874 is an estimate of a \( p \)-value, but it’s an underestimate.

**Q:** Is there any alternate hypothesis test that would have rendered a judgment of success?

**A:** In theory, yes. For example, the hypothesis test that always yields a failure except when \((i, j) = (68, 67)\) would certainly keep false positives below 5%, and would still have granted a success in this instance.

But if we restrict our search to confidence intervals and hypothesis tests found in the literature or in commercial statistics software, the answer seems to be no. With every one of 15 other methods, the results didn’t achieve statistical significance. Thus, the ATLAS study’s apparent success at meeting the primary end-point was solely an artifact of the Wald interval’s consistent undercoverage.

<table>
<thead>
<tr>
<th>Method</th>
<th>( p )-value or conf. bound</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional ( z )-test (Boston Scientific’s method)</td>
<td>( p = 0.04874 )</td>
<td>Pass</td>
</tr>
<tr>
<td>( z )-test, constrained max likelihood standard error</td>
<td>( p = 0.05151 )</td>
<td>Fail</td>
</tr>
<tr>
<td>( z )-test with Yates continuity correction</td>
<td>( c = 0.03095 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Agresti-Caffo ( I_4 ) interval</td>
<td>( p = 0.05021 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Wilson score</td>
<td>( c = 0.03015 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Wilson score with continuity correction</td>
<td>( c = 0.03094 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Farrington &amp; Manning score</td>
<td>( p = 0.05151 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Miettinen &amp; Nurminen score</td>
<td>( p = 0.05156 )</td>
<td>Fail</td>
</tr>
<tr>
<td>Gart &amp; Nam score</td>
<td>( p = 0.05096 )</td>
<td>Fail</td>
</tr>
<tr>
<td>NCSS’s bootstrap method</td>
<td>( c = 0.03006 )</td>
<td>Fail</td>
</tr>
<tr>
<td>NCSS’s quasi-exact Chen</td>
<td>( c = 0.03016 )</td>
<td>Fail</td>
</tr>
<tr>
<td>NCSS’s exact double-binomial test</td>
<td>( p = 0.05470 )</td>
<td>Fail</td>
</tr>
<tr>
<td>StatXact’s approximate unconditional test of non-inferiority</td>
<td>( p = 0.05151 )</td>
<td>Fail</td>
</tr>
<tr>
<td>StatXact’s exact unconditional test of non-inferiority</td>
<td>( p = 0.05138 )</td>
<td>Fail</td>
</tr>
<tr>
<td>StatXact’s exact CI based on difference of observed rates</td>
<td>( c = 0.03737 )</td>
<td>Fail</td>
</tr>
<tr>
<td>StatXact’s approximate CI from inverted 2-sided test</td>
<td>( c = 0.03019 )</td>
<td>Fail</td>
</tr>
<tr>
<td>StatXact’s exact CI from inverted 2-sided test</td>
<td>( c = 0.03032 )</td>
<td>Fail</td>
</tr>
</tbody>
</table>

**Q:** \( N = 855 \) for Liberte and \( N = 956 \) for Express, and the observed counts were 68 and 67. \( np \) and \( n(1-p) \) are almost certainly greater than 5. Isn’t that more than enough to justify use of a normal-distribution approximation?

**A:** Not all normal approximations are created equal. The rules of thumb that apply to “superiority” trials — attempts to reject the null hypothesis that one rate is equal to another — may not be valid for “non-inferiority” trials.

Consider the following comparison of a standard \( z \)-test for superiority (in red, with a pooled estimate of the variance, since the null hypothesis is of equal rates), to the ATLAS trial’s unpooled \( z \)-test, in blue:
We can say that the Wald interval consistently undercovers (provides less than 95% coverage) for any pair of plausible TVR rates on the critical line at the boundary of the null hypothesis. Even off the critical line, the Wald interval consistently undercovers at every possible pair of rates within a 95% confidence region around the observed rates:

Do the rules of thumb developed for superiority testing apply to non-inferiority?

When testing superiority, the z-test's coverage is close to nominal, and test rarely provides less than 95% coverage...

...but for non-inferiority, the error is larger and in the wrong direction. The result is a test that doesn't bound Type I error or provide nominal coverage anywhere along the parameter region of interest.

z-test for superiority along $p_1 = p_2$

z-test for non-inferiority with 0.03 margin along $p_1 + 0.03 = p_2$
Q: I’m a Bayesian. What is the a posteriori probability that Liberte is markedly inferior to Express, meaning Liberte’s rate of TVR after 9 months is at least 3 percentage points higher than Express’s?

A: We can calculate the a posteriori probability that the inferiority hypothesis is true, given a uniform prior and the observed outcome.

Assume the underlying TVR rates are chosen uniformly between 0% and 100%. In the Liberte group, we measured 68 TVRs and 787 non-TVRs. In the Express group, it was 67 TVRs and 889 non-TVRs. Then the probability density function for the outcome \((x, y) = (\pi_{TVRe}, \pi_{TVRl})\) is given by 

\[
 f(x; 68, 890)f(y; 69, 788),
\]

where \(f(x; \alpha, \beta)\) is the PDF of the beta distribution:

\[
 f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}
\]

The probability that \(\pi_{TVRl} - \pi_{TVRe} > 0.03\) is given by integrating over that region of the parameter space:

\[
 \int_0^1 \int_{\min(x+0.03,1)}^1 f(x; 68, 890)f(y; 69, 788) \, dy \, dx \approx 0.050737979 \ldots
\]

Thus, given the outcome of the ATLAS trial and the assumption that the TVR rates were drawn uniformly between 0 and 1, the probability that \(\pi_{TVRl} - \pi_{TVRe} > 0.03\) is 5.074%.

If we retain the traditional 5% inference threshold, then we cannot reject the hypothesis that Liberte is markedly inferior to Express.

Q: What do statistics textbooks say about results with \(p\)-values close to 0.05?

A: In general, they caution that it’s dangerous to rely on an approximation to claim a borderline finding is on one side of a line or another. Fleiss et al., Statistical Methods for Rates and Proportions (3rd ed.), p. 62 says:

> In situations where the choice of test statistic or confidence interval would make a difference to the inferences drawn, exact methods should be relied upon rather than normal approximations.

> In that way disagreements over “borderline significant” results can be minimized.

Unlike using a normal-distribution approximation, exact methods guarantee that false positives will be controlled at the desired rate.

Q: Even though the “entry-criteria-matched” Express patients fell within the same outer limits as the Liberte patients, the characteristics of the two groups’ blockages varied significantly. On average, the Liberte patients had more complex lesions than the Express patients. What happens if you adjust for those differences?

A: Boston Scientific has done this analysis internally. After adjustment for differences that were measured, the 95% one-sided confidence bound on the risk difference is roughly 2.6 percentage points (Liberte rate higher than Express) — within the 3-percentage-point definition of inferiority. However, such an adjustment cannot correct for the fact that Express patients, their doctors, and the event adjudicators were blinded as to which stent was used (Taxus Express vs. bare-metal Express), whereas the Liberte patients, doctors, and adjudicators were aware of the device’s identity. Also, the Liberte patients received their stents up to three years after the Express patients, meaning doctors may have been more experienced or practices may have differed.