Bayesianness and Frequentism

Keith Winstein

keithw@mit.edu

October 13, 2009

Axioms of Probability

Let S be a finite set called the *sample space*, and let A be any subset of S, called an *event*. The *probability* $P(A)$ is a real-valued function that satisfies:

- $P(A) \geq 0$
- $P(S) = 1$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

For infinite sample space, third axiom is that for an infinite sequence of disjoint subsets $A_1, A_2, \ldots,$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

.

Some Theorems

- $P(\overline{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A) \leq P(B)$ if $A \subset B$
- $P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$

Joint & Conditional Probability

- If A and B are two events (subsets of S), then call $P(A \cap B)$ the *joint probability* of A and B.
- Define the *conditional probability of A given B* as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

.

- A and B are said to be *independent* if $P(A \cap B) = P(A)P(B)$.
- If A and B are independent, then $P(A|B) = P(A)$.

Bayes' Rule

We have:

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$

Therefore:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

And Bayes' Rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

On the islands of Ste. Frequentiste and Bayesienne...

On the islands of Ste. Frequentiste and Bayesienne...

The king has been poisoned!

On the islands of Ste. Frequentiste and Bayesienne...

The king of Ste. F & B has been poisoned! It's a conspiracy. An order goes out to the regional governors of Ste. Frequentiste and of Isle Bayesienne: find those responsible, and jail them.

Dear Governor: Attached is a blood test for proximity to the poison that killed the king. It has a 0% rate of false negative and a 1% rate of false positive. Administer it to everybody on your island, and if you conclude they're guilty, jail them.

BUT REMEMBER THE NATIONWIDE LAW: **We must be 95% certain of guilt to send a citizen to jail.**

On Ste. Frequentiste:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

- $P(E^+|\text{Guilty}) = 1$
- $P(E^-|\text{Guilty}) = 0$
- $P(E^+|\text{Innocent}) = 0.01$
- $P(E^-|\text{Innocent}) = 0.99$

How to interpret the law?

"We must be 95% certain of guilt" $\Rightarrow$

On Ste. Frequentiste:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

- $P(E^+|\textsc{Guilty}) = 1$
- $P(E^-|\textsc{Guilty}) = 0$
- $P(E^+|\textsc{Innocent}) = 0.01$
- $P(E^-|\textsc{Innocent}) = 0.99$

How to interpret the law?

"We must be 95% certain of guilt" $\Rightarrow P(\textsc{Jail}|\textsc{Innocent}) \leq 5\%$.

On Ste. Frequentiste:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

- $P(E^+|\textsc{Guilty}) = 1$
- $P(E^-|\textsc{Guilty}) = 0$
- $P(E^+|\textsc{Innocent}) = 0.01$
- $P(E^-|\textsc{Innocent}) = 0.99$

How to interpret the law?
"We must be 95% certain of guilt" $\Rightarrow P(\textsc{Jail}|\textsc{Innocent}) \leq 5\%$.

Governor F.: *Ok, what if I jail everybody with a positive test result? Then* $P(\textsc{Jail}|\textsc{Innocent}) = P(E^+|\textsc{Innocent}) = 1\%$. *That's less than 5%, so we're obeying the law."*

On Isle Bayesienne:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

How to interpret the law?

"We must be 95% certain of guilt" $\Rightarrow$

On Isle Bayesienne:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

How to interpret the law?
"We must be 95% certain of guilt" $\Rightarrow P(\textsc{Innocent}|\textsc{Jail}) \leq 5\%$.

Governor B.: *Can I jail everyone with a positive result? I'll apply Bayes' rule...*

$$P(\textsc{Innocent}|E^+) = P(E^+|\textsc{Innocent})\frac{P(\textsc{Innocent})}{P(E^+)}$$

We need to know $P(\textsc{Innocent})$.

On Isle Bayesienne:

The test has a 0% rate of false negative and a 1% rate of false positive. **We must be 95% certain of guilt to send a citizen to jail.**

How to interpret the law?
"We must be 95% certain of guilt" $\Rightarrow P(\textsc{Innocent}|\textsc{Jail}) \leq 5\%$.

Governor B.: *Can I jail everyone with a positive result? I'll apply Bayes' rule...*

$$P(\textsc{Innocent}|E^+) = P(E^+|\textsc{Innocent})\frac{P(\textsc{Innocent})}{P(E^+)}$$

We need to know $P(\textsc{Innocent})$. Governor B.: *Hmm, I will assume that 10% of my subjects were guilty of the conspiracy.* $P(\textsc{Innocent}) = 0.9$.

On Isle Bayesienne:

Apply Bayes' rule

- We know the conditional probabilities of the form $P(E^+|\textsc{Guilty})$.
- Governor knows the "overall" probability of each event $\textsc{Guilty}$ and $\textsc{Innocent}$. Since this is our estimate of the chance someone is guilty *before* a blood test, we call it the *prior probability*.
- We can combine prior and conditional probabilities to form the joint probability matrix of the form $P(E^+ \cap \textsc{Guilty})$.
- Then, turn the joint probabilities into conditiononal probabilities, e.g., $P(\textsc{Guilty}|E^+)$.
- Result: $P(\textsc{Innocent}|E^+) \approx 8\%$. Too high!

On the islands of Ste. Frequentiste and Bayesienne...

Results:

- More than 1% of Ste. Frequentiste goes to jail.
- On Isle Bayesienne, 10% are guilty, but nobody goes to jail.
- The disagreement isn't about math. It isn't necessarily about philosophy. Here, the frequentist and Bayesian used tests that met different constraints and got different results.

The Constraints

- The frequentist cares about the rate of jailings among innocent people and wants it to be less than 5%. Concern: **overall rate of false positive**.
- The Bayesian cares about the rate of innocence among jail inmates and wants it to be less than 5%. Concern: **rate of error among positives**.
- The Bayesian had to make assumptions about the overall, or prior, probabilities.

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

It can be proven that most claimed research findings are false.

yet ill-founded strategy of claiming conclusive research findings solely on

is characteristic of the vary a lot depending o field targets highly like or searches for only on true relationships amo and millions of hypoth be postulated. Let us a for computational simp circumscribed fields wl is only one true relatio many that can be hypo the power is similar to

Why Most Published Research Findings Are False, Ioannidis JPA, PLoS Medicine Vol. 2, No. 8, e124
doi:10.1371/journal.pmed.0020124

Confidence & Credibility

- For similar reasons, frequentists and Bayesians express uncertainty differently.
- Both use *intervals*: a function that maps each possible observation to a set of parameters.
- Frequentists use *confidence intervals*. For every value of the parameter, the *coverage* is the probability that the interval will include that value. The *confidence parameter* is formally the minimum of the coverage.
- Bayesians use *credible (or credibility) intervals*. For every outcome, the interval gives a set of parameters whose conditional probability sums to at least the specified credibility. Needs a prior.

Confidence & Credibility

- Confidence interval: *"Even before we start, we can promise that the probability the experiment will produce a wrong answer in the end is less than 5% — just like the probability that Ste. Frequentist will jail an innocent person. Our confidence interval might sometimes be nonsense, but as long as that happens less than 5% of the time, it's ok."*

- Credibility interval: *"Now that we took data, we can say that the true value lies within this interval with 95% probability. This required an assumption of the overall probability of each parameter value. If God punishes us by choosing an unlikely value of the parameter, our credible interval could be very misleading."* (Billion to one example.)

A Pathological Example

Cookie jars **A**, **B**, **C**, **D** have the following distribution of cookies with chocolate chips:

$P(\ chips \mid jar\ )$	**A**	**B**	**C**	**D**
0	1	17	14	27
1	1	20	22	70
2	70	22	20	1
3	28	20	22	1
4	0	21	22	1
total	100%	100%	100%	100%

Let's construct a **70%** confidence interval.

70% Confidence Intervals

Cookie jars **A**, **B**, **C**, **D** have the following distribution of cookies with chocolate chips:

$P(\ chips \mid jar\ )$	**A**	**B**	**C**	**D**
0	1	17	14	27
1	1	**[20**	**22**	**70]**
2	**[70**	**22**	**20]**	1
3	28	**[20**	**22]**	1
4	0	**[21**	**22]**	1
coverage	70%	83%	86%	70%

The **70%** confidence interval has at least 70% coverage for every value of the parameter.
Now assume a uniform prior and calculate $P(\ jar \cap\ chips\ )$.

Joint Probabilities

Cookie jars **A**, **B**, **C**, **D** have equal chance of being selected, and the following joint distribution of jar and chips:

$P(\ jar \cap chips\ )$	**A**	**B**	**C**	**D**	total
0	1/4	17/4	14/4	27/4	14.75%
1	1/4	20/4	22/4	70/4	28.25%
2	70/4	22/4	20/4	1/4	28.25%
3	28/4	20/4	22/4	1/4	17.75%
4	0/4	21/4	22/4	1/4	11.00%
total	25%	25%	25%	25%	

Now calculate $P(\ jar \mid chips\ )$.

$P(\ outcome\ |\theta)$

Cookie jars **A**, **B**, **C**, **D** have the following conditional probability of each jar given the number of chips:

$P(\ jar \mid chips\ )$	**A**	**B**	**C**	**D**	total
0	1.7	28.8	23.7	45.8	100%
1	0.9	17.7	19.5	61.9	100%
2	61.9	19.5	17.7	0.9	100%
3	39.4	28.2	31.0	1.4	100%
4	0.0	47.7	50.0	2.3	100%

Now let's make **70%** credibility intervals.

70% Credibility Intervals

Cookie jars **A**, **B**, **C**, **D** have the following conditional probability of each jar given the number of chips:

$P($ *jar* $\mid$ *chips* $)$	**A**	**B**	**C**	**D**	credibility
0	1.7	**[28.8]**	23.7	**[45.8]**	75%
1	0.9	17.7	**[19.5**	**61.9]**	81%
2	**[61.9**	**19.5]**	17.7	0.9	81%
3	**[39.4]**	28.2	**[31.0]**	1.4	70%
4	0.0	**[47.7**	**50.0]**	2.3	98%

Confidence & Credible Intervals

$4P(\ jar \cap chips\ )$	**A**	**B**	**C**	**D**	credibility
0	1	17	14	27	**0%**
1	1	**[20**	**22**	**70]**	99%
2	**[70**	**22**	**20]**	1	99%
3	28	**[20**	**22]**	1	**59%**
4	0	**[21**	**22]**	1	98%
coverage	70%	83%	86%	70%	

$4P(\ jar \cap chips\ )$	**A**	**B**	**C**	**D**	credibility
0	1	**[17]**	14	**[27]**	75%
1	1	20	**[22**	**70]**	81%
2	**[70**	**22]**	20	1	81%
3	**[28]**	20	**[22]**	1	70%
4	0	**[21**	**22]**	1	98%
coverage	98%	**60%**	**66%**	97%	

The TAXUS ATLAS Experiment

- Data: 1,811 people in one of two groups.
- 956 people are assigned to CONTROL and 855 people to TREATMENT.
- We're counting bad events in each group.
- We want to know: comparing proportions of patients who get an event, is TREATMENT **non-inferior** to CONTROL, with a three-percentage-point margin, at the $p < 0.05$ level?
  - CONTROL 7% vs. TREATMENT 10.5% would be "inferior."
  - CONTROL 7% vs. TREATMENT 9.5% would be "non-inferior."
- We assume each population has a certain true rate of events, π_t and π_c.
- We record the number of patients who get an event in our experiment, n_t and n_c.
- Is there 95% confidence that $\pi_t - \pi_c < 0.03$?

ATLAS Trial Solution

- Use a one-sided 95% confidence interval for $\pi_t - \pi_c$. If its upper limit is less than 0.03, accept. Otherwise reject.
- Confidence interval: approximate *each binomial separately* with a normal distribution. Known as Wald interval.
- If we sample a Bernoulli trial N times and get i successes, we can approximate source distribution as a Gaussian with mean i/N and variance $\frac{i(N-i)}{N^3}$.
- Calculate the distribution of the difference of these two binomials, and see if 95% of the area is less than 0.03.
- $$p \approx \text{area} = \int_{0.03}^{\infty} \mathcal{N}\left(\frac{i}{m} - \frac{j}{n}, \frac{i(m-i)}{m^3} + \frac{j(n-j)}{n^3}\right)$$

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

ATLAS Results

- We measure 68/855 events in $\textsc{Treatment}$ (7.95%), and 67/956 events in $\textsc{Control}$ (7.01%).
- Procedure: if $\mathrm{area} < 5\%$, we accept. Area is serving the function of a *p-value*: an upper bound on the rate of false positives we're willing to accept. If our tolerance were 1%, cutoff would be 0.01.
- $p \approx \int_{0.03}^{\infty} \mathcal{N}\left(\frac{i}{m} - \frac{j}{n}, \frac{i(m-i)}{m^3} + \frac{j(n-j)}{n^3}\right) = 0.0487395\ldots$.
- Accept.

ATLAS Results (May 2006)

TAXUS ATLAS Trial Supports Superior Deliverability and Proven Outcomes of TAXUS(R) Liberte(TM) Stent System; Boston Scientific's second generation stent compares favorably to market leading TAXUS Express2(TM) stent system, even with more complex lesions

May 16, 2006 — NATICK, Mass. and PARIS, May 16 /PRNewswire-FirstCall/ — Boston Scientific Corporation today announced nine-month data from its TAXUS ATLAS clinical trial. The results confirmed safety and efficacy and demonstrated the superior deliverability of the TAXUS(R) Liberte(TM) paclitaxel-eluting stent system compared to the TAXUS Express2(TM) paclitaxel-eluting stent system. [...] **The trial met its primary endpoint** of nine-month target vessel revascularization (TVR), a measure of the effectiveness of a coronary stent in reducing the need for a repeat procedure.

ATLAS Results (April 2007)

Turco et al., *Polymer-Based, Paclitaxel-Eluting TAXUS Liberté Stent in De Novo Lesions*, Journal of the American College of Cardiology, Vol. 49, No. 16, 2007.

Results: The primary non-inferiority end point was met with the 1-sided 95% confidence bound of 2.98% less than the pre-specified non-inferiority margin of 3% ($p = 0.0487$).

Statistical methodology. P values are 2-sided unless specified otherwise. Student t test was used to compare independent continuous variables, while chi-square or Fisher exact test was used to compare proportions.

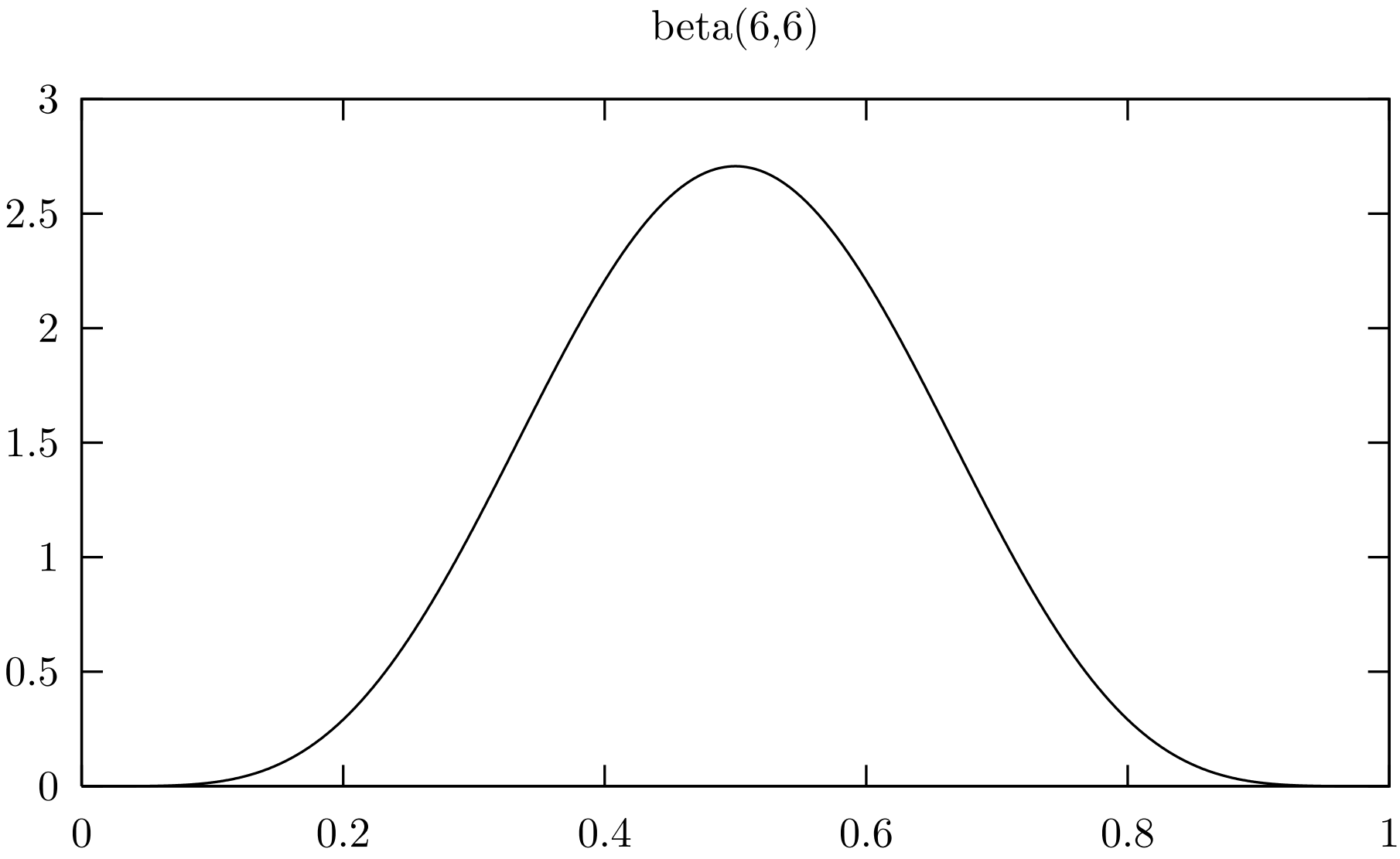
Bayesian Results

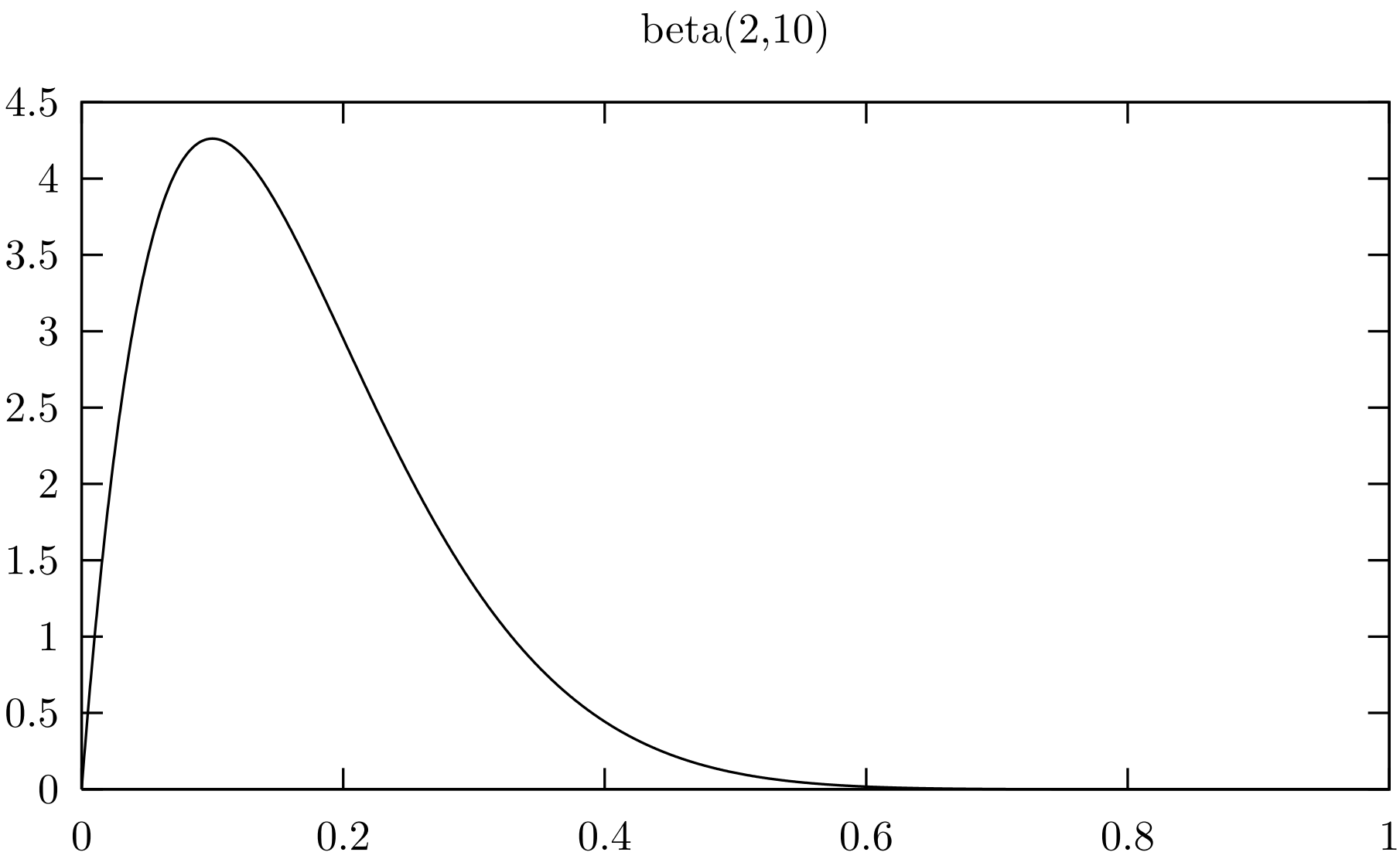
- Bayesian says, "Let's assume I know nothing about π_t and π_c *a priori*. I assume God chose them randomly on [0,1], independently and with uniform probability."
- Then we sample each binomial: in TREATMENT, we do 855 samples and get 68 heads. In CONTROL, we do 956 samples and get 67 heads.
- For a particular π_t and 855 samples, probability of k heads is $\mathrm{Bin}(x; 855, \pi_t)$.
- $$\mathrm{Bin}(k; N, \pi) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$$
- Apply Bayes' rule.

Bayesian Results

- Likelihood: $L_{Nk}(\pi) = \binom{N}{k}\pi^k(1-\pi)^{N-k}$
- Probability: Apply Bayes' rule. With a uniform prior, just normalize. Result is called a Beta distribution.
- $$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

  where $\alpha =$ heads observed plus one, and $\beta =$ tails observed plus one.


beta(6,6)
3
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

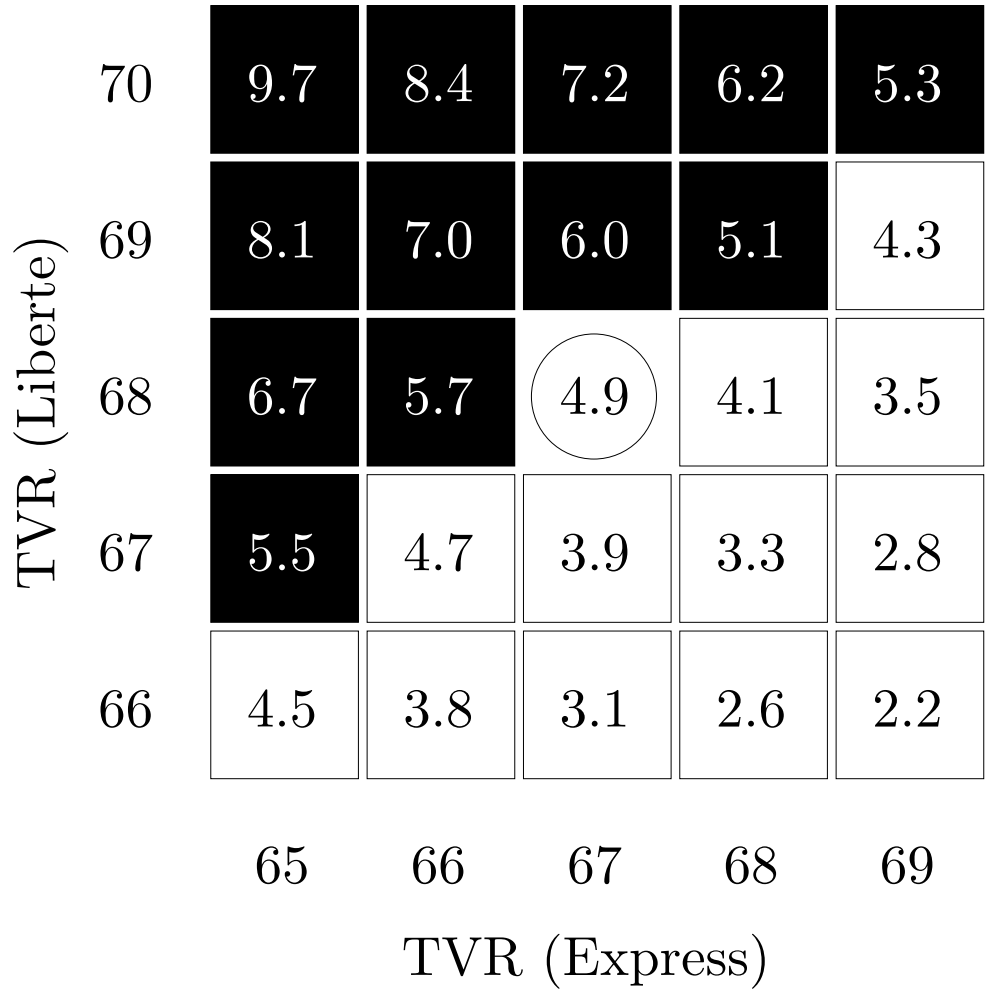

beta(2,10)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

Bayesian Results

- We got 68 heads and 787 tails in TREATMENT, and 67 heads and 889 tails in CONTROL. With a uniform prior, we calculate the *a posteriori* probability of each π.
- $\pi_c \sim \beta(x; 68, 890)$
- $\pi_t \sim \beta(x; 69, 788)$
- What's the *a posteriori* probability that $\pi_t - \pi_c < 0.03$?
- $$\int_0^1 \int_{\min(x+0.03,1)}^1 \beta(x; 68, 890)\beta(y; 69, 788)\, dy\, dx \approx 0.050737979\ldots$$
- We think the probability is more than 5%.

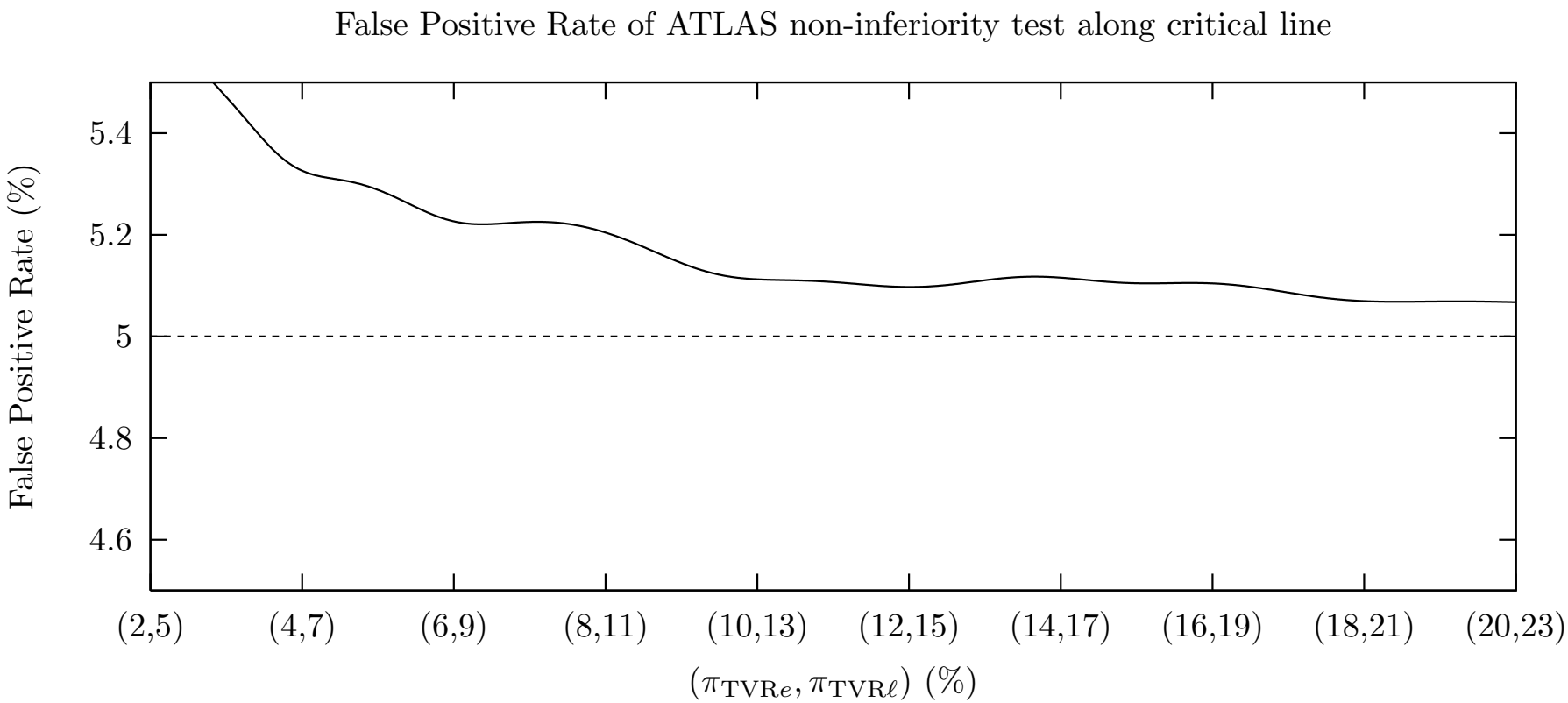
The Ultimate Close Call

Wald's area ($\approx p$) with $(m, n) = (855, 956)$



The Wald Interval Undercovers

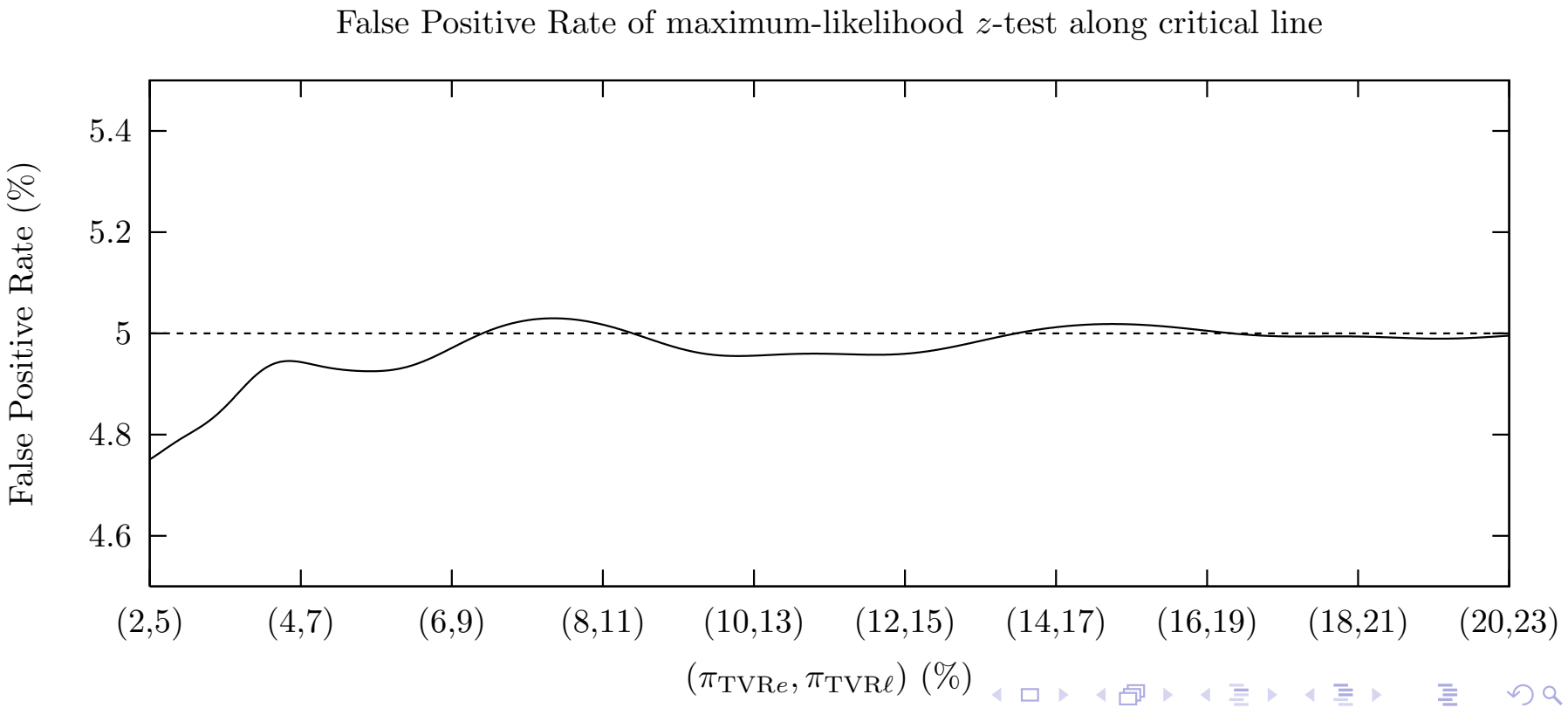
Is this a disagreement between frequentist and Bayesian methods? In this case, no. Our confidence interval doesn't have 95% coverage, so the test didn't bound the rate of false positives by 0.05. The approximation is lousy in this context.



One solution: constrained variance

The Wald interval approximated each binomial *separately* as a Gaussian, with variance of $\frac{i(N-i)}{N^3}$. (E.g., 7% and 8%.) But this is not consistent with H_0, which says $\pi_t > \pi_c + 0.03$.

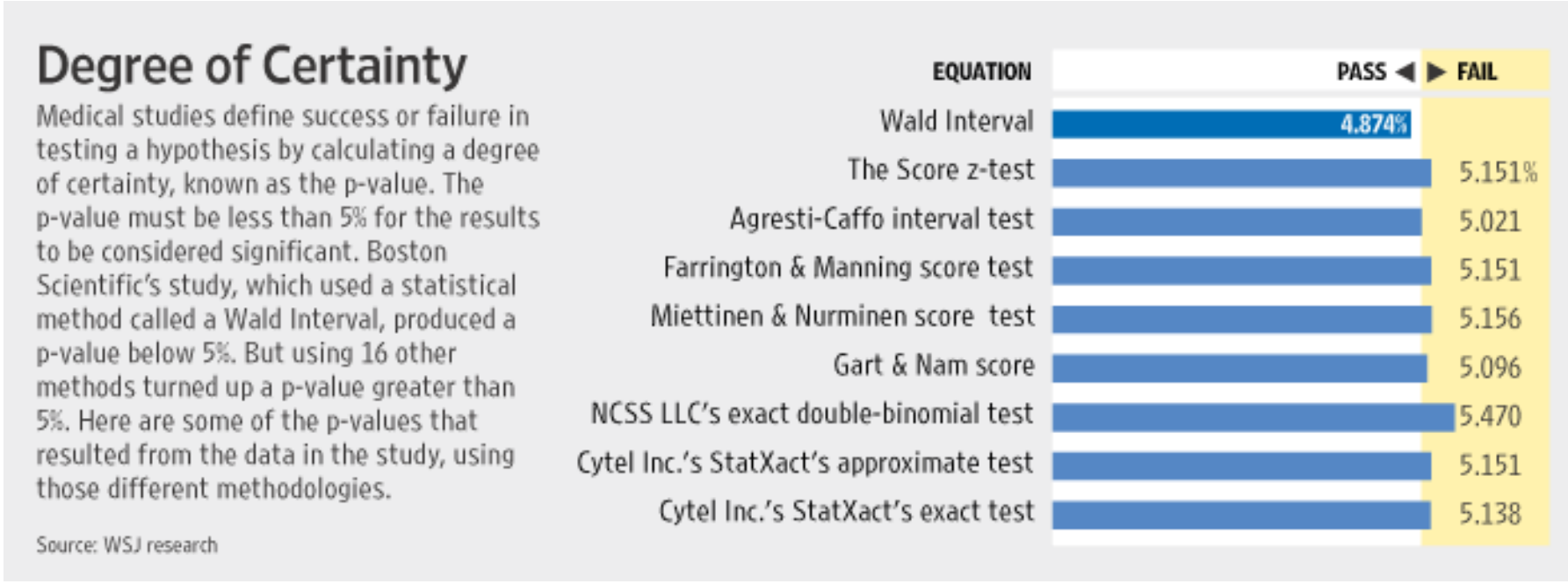
One improvement is to approximate the variances by finding the most likely pair consistent with H_0 (i.e., separated by 3 percentage points). E.g., 6% and 9%.



Every other published interval fails to exclude inferiority.

Method	***p*-value or confidence bound**	**Result**
Wald interval	$p = 0.04874$	**Pass**
z-test, constrained max likelihood standard error	$p = 0.05151$	Fail
z-test with Yates continuity correction	$c = 0.03095$	Fail
Agresti-Caffo I_4 interval	$p = 0.05021$	Fail
Wilson score	$c = 0.03015$	Fail
Wilson score with continuity correction	$c = 0.03094$	Fail
Farrington & Manning score	$p = 0.05151$	Fail
Miettinen & Nurminen score	$p = 0.05156$	Fail
Gart & Nam score	$p = 0.05096$	Fail
NCSS's bootstrap method	$c = 0.03006$	Fail
NCSS's quasi-exact Chen	$c = 0.03016$	Fail
NCSS's exact double-binomial test	$p = 0.05470$	Fail
StatXact's approximate unconditional test of non-inferiority	$p = 0.05151$	Fail
StatXact's exact unconditional test of non-inferiority	$p = 0.05138$	Fail
StatXact's exact CI based on difference of observed rates	$c = 0.03737$	Fail
StatXact's approximate CI from inverted 2-sided test	$c = 0.03019$	Fail
StatXact's exact CI from inverted 2-sided test	$c = 0.03032$	Fail

Nerdiest chart contender?



Boston Scientific Stent Study Flawed

By Keith J. Winstein

A heart stent manufactured by Boston Scientific Corp. and expecting approval for U.S. sales is backed by flawed research despite the company's claims of success in a clinical trial, according to a Wall Street Journal review of the data.

Boston Scientific submitted the results of the 2006 trial to the Food and Drug Administration to gain U.S. approval for the Taxus Liberte, which already is one of the top-selling stents abroad. Coronary stents—tiny scaffolds that prop open arteries clogged by heart disease—are one of the most popular methods for treating heart patients, and have been implanted in more than 15 million people world-wide.

But Boston Scientific's claim was based on a flawed statistical equation that favored the Liberte stent, a Journal analysis has found. Using a number of other methods of calculation—including 14 available in off-the-shelf software programs—the Liberte study would have been a failure by the common standards of statistical significance in research.

Boston Scientific isn't the only company to use the equation, known as a Wald interval, which has long been criticized by statisticians for exaggerating the certainty of research results. Rivals Medtronic Inc. and Abbott Laboratories have used the same equation in stent studies.

Boston Scientific

Boston Scientific is seeking FDA approval for its **Taxus Liberte stent**.

But in those cases, any boost provided by the Wald equation wouldn't have changed the outcome of the study. In the Liberte study, the equation's shortcomings meant the difference between success and failure in the study's main goal.

The difference also sheds light on the leeway that device makers have when designing studies for the FDA. Studies designed to satisfy the requirements of the FDA's medical-device branch can be less rigorous than those aimed at winning U.S. approval for drugs. That is partly because of a 1997 federal law aimed at lessening the regulatory requirements on device makers.

The FDA declined to specifically discuss its deliberations of the Liberte, which is still under review by the agency.

Boston Scientific doesn't agree that it made a mistake or that the study failed to reach statistical significance. "We used standard methodology that we discussed with the FDA up front, and then executed," said Donald Baim, Boston Scientific's chief scientific and medical officer.

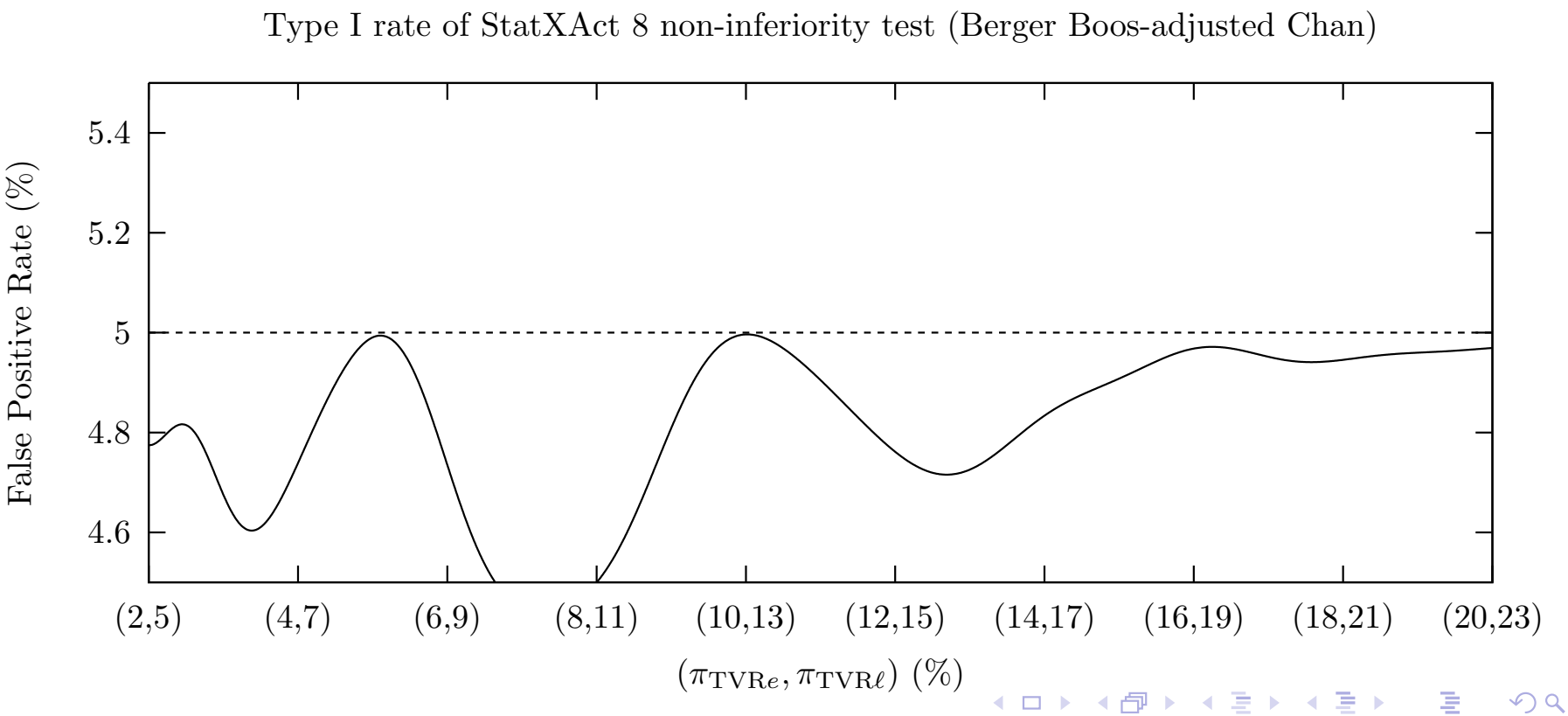
Please turn to page B6

World's most advanced non-inferiority test

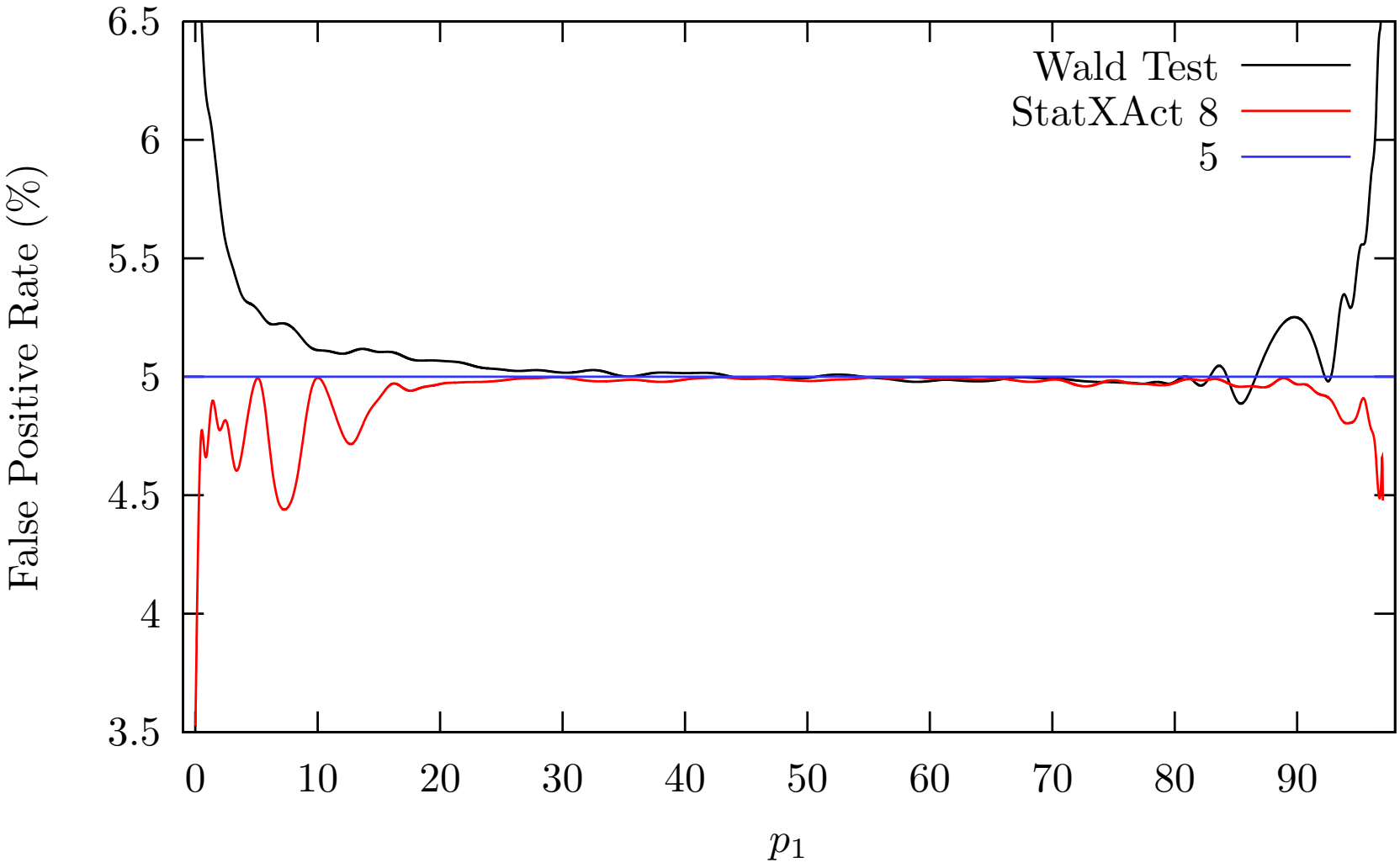
The StatXAct 8 software package sells for $1,000 and takes 15 minutes to calculate a single p-value. (Mention very nice lunch.)
"Other statistical applications often rely on large-scale assumptions for inferences, risking incorrect conclusions from data sets not normally distributed. StatXact utilizes Cytel's own powerful algorithms to make exact inferences by permuting the actually observed data, eliminating the need for distributional assumptions."

World's most advanced non-inferiority test

The StatXAct 8 software package sells for $1,000 and takes 15 minutes to calculate a single *p*-value. (Mention very nice lunch.)
"Other statistical applications often rely on large-scale assumptions for inferences, risking incorrect conclusions from data sets not normally distributed. StatXact utilizes Cytel's own powerful algorithms to make exact inferences by permuting the actually observed data, eliminating the need for distributional assumptions."



Both tests, together



Pre-specification

- To meet the frequentist's constraint, every detail of the experiment and testing procedure has to be *pre-specified*.
- Two different tests may each have a false positive rate less than 5%. But if you can pick which test to use after the fact, you'll get a false positive rate more than 5%. The reason: the *union* of the two tests, although each would be valid by itself, doesn't have a false positive rate less than 5%.
- Not so for the Bayesian. Posterior probability is determined by the prior and the design of experiment. Bayesian constraint isn't violated by switching priors after the fact.
- Blinding through the analysis is still a good idea.

Final Thoughts

- What's important: say what your criteria are, and make sure the test or interval meets them.
- Don't be surprised if frequentist and Bayesian approaches differ in their results.
- Sometimes they will agree numerically but not on what the numbers mean!
- If they disagree starkly, you have bigger problems than your interpretation of probability.
- Same goes if the Bayesian answer depends heavily on the prior. If two reasonable priors give starkly disagreeing results, you don't have a good answer.