

And the Bayesians and the frequentists
shall lie down together...

Keith Winstein

MIT CSAIL

February 12, 2014



Axioms of Probability (1933)

S : a finite set (the **sample space**)

A : any subset of S (an **event**)

$P(A)$: the **probability** of A satisfies

- ▶ $P(A) \in \mathbb{R}$
- ▶ $P(A) \geq 0$
- ▶ $P(S) = 1$
- ▶ $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

If S infinite, axiom becomes: for an infinite sequence of disjoint subsets A_1, A_2, \dots ,

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

Some Theorems

- ▶ $P(\bar{A}) = 1 - P(A)$
- ▶ $P(\emptyset) = 0$
- ▶ $P(A) \leq P(B)$ 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

- ▶ For $A, B \subseteq S$, $P(A \cap B)$ is **joint probability** of A and B .
- ▶ The **conditional probability** of A given B in:

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

- ▶ A and B are **independent** iff $P(A \cap B) = P(A)P(B)$.
- ▶ A, B independent $\rightarrow P(A|B) = P(A)$.

Bayes' Theorem

We have:

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

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

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

Bayes' Theorem:

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

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

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



The king has been poisoned!

A letter goes out. . .

Dear Governor: Attached is a blood test for proximity to the poison. **It has a 0% rate of false negative and a 1% rate of false positive.** Jail those responsible.

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

On Ste. Frequentiste:

Test has a 0% rate of false negative and a 1% rate of false positive.
You must be 95% certain to send a citizen to jail.

- ▶ $P(\textit{Positive} \mid \textit{GUILTY}) = 1$
- ▶ $P(\textit{Negative} \mid \textit{GUILTY}) = 0$
- ▶ $P(\textit{Positive} \mid \textit{INNOCENT}) = 0.01$
- ▶ $P(\textit{Negative} \mid \textit{INNOCENT}) = 0.99$

How to interpret the law?

“We must be 95% certain” →

On Ste. Frequentiste:

Test has a 0% rate of false negative and a 1% rate of false positive.
You must be 95% certain to send a citizen to jail.

- ▶ $P(\text{Positive} \mid \text{GUILTY}) = 1$
- ▶ $P(\text{Negative} \mid \text{GUILTY}) = 0$
- ▶ $P(\text{Positive} \mid \text{INNOCENT}) = 0.01$
- ▶ $P(\text{Negative} \mid \text{INNOCENT}) = 0.99$

How to interpret the law?

“We must be 95% certain” → $P(\text{Jail} \mid \text{INNOCENT}) \leq 0.05$

On Ste. Frequentiste:

Test has a 0% rate of false negative and a 1% rate of false positive.
You must be 95% certain to send a citizen to jail.

- ▶ $P(\text{Positive} \mid \text{GUILTY}) = 1$
- ▶ $P(\text{Negative} \mid \text{GUILTY}) = 0$
- ▶ $P(\text{Positive} \mid \text{INNOCENT}) = 0.01$
- ▶ $P(\text{Negative} \mid \text{INNOCENT}) = 0.99$

How to interpret the law?

“We must be 95% certain” → $P(\text{Jail} \mid \text{INNOCENT}) \leq 0.05$

Can Positive → Jail? Yes.

On Isle Bayesienne:

Test has a 0% rate of false negative and a 1% rate of false positive.
You must be 95% certain to send a citizen to jail.

- ▶ $P(\textit{Positive} \mid \textit{GUILTY}) = 1$
- ▶ $P(\textit{Negative} \mid \textit{GUILTY}) = 0$
- ▶ $P(\textit{Positive} \mid \textit{INNOCENT}) = 0.01$
- ▶ $P(\textit{Negative} \mid \textit{INNOCENT}) = 0.99$

How to interpret the law?

“We must be 95% certain” →

On Isle Bayesienne:

Test has a 0% rate of false negative and a 1% rate of false positive.
You must be 95% certain to send a citizen to jail.

- ▶ $P(\textit{Positive} \mid \textit{GUILTY}) = 1$
- ▶ $P(\textit{Negative} \mid \textit{GUILTY}) = 0$
- ▶ $P(\textit{Positive} \mid \textit{INNOCENT}) = 0.01$
- ▶ $P(\textit{Negative} \mid \textit{INNOCENT}) = 0.99$

How to interpret the law?

“We must be 95% certain” → $\mathbf{P(\textit{INNOCENT} \mid \textit{Jail}) \leq 0.05}$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Positive}) = \frac{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Positive}) = \frac{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow P(\text{INNOCENT} \mid \text{Jail}) \leq 0.05$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} \mid \text{Jail}) = \frac{P(\text{Positive} \mid \text{INNOCENT}) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow P(\text{INNOCENT} \mid \text{Jail}) \leq 0.05$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} \mid \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT}) + P(\text{Positive} | \text{GUILTY}) P(\text{GUILTY})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT}) + P(\text{Positive} | \text{GUILTY}) P(\text{GUILTY})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow P(\text{INNOCENT} | \text{Jail}) \leq 0.05$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{P(\text{Positive} | \text{INNOCENT}) P(\text{INNOCENT}) + P(\text{Positive} | \text{GUILTY}) P(\text{GUILTY})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} \mid \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} \mid \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{1 - (0.99) P(\text{INNOCENT})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{1 - (0.99) P(\text{INNOCENT})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” → $P(\text{INNOCENT} \mid \text{Jail}) \leq 0.05$
- ▶ Can Positive → Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} \mid \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{1 - (0.99) P(\text{INNOCENT})}$$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” → $P(\text{INNOCENT} \mid \text{Jail}) \leq 0.05$
- ▶ Can Positive → Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} \mid \text{Jail}) = \frac{(0.01) P(\text{INNOCENT})}{1 - (0.99) P(\text{INNOCENT})}$$

- ▶ $P(\text{Innocent}) = ???$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) \cdot P(\text{INNOCENT})}{1 - (0.99) \cdot P(\text{INNOCENT})}$$

- ▶ $\mathbf{P(\text{Innocent})} = 0.9 \rightarrow$

Isle Bayesienne: the need for prior assumptions

- ▶ “We must be 95% certain” $\rightarrow \mathbf{P(\text{INNOCENT} | \text{Jail})} \leq \mathbf{0.05}$
- ▶ Can Positive \rightarrow Jail?
- ▶ Apply Bayes' theorem

$$P(\text{INNOCENT} | \text{Jail}) = \frac{(0.01) \cdot P(\text{INNOCENT})}{1 - (0.99) \cdot P(\text{INNOCENT})}$$

- ▶ $\mathbf{P(\text{Innocent})} = 0.9 \rightarrow \mathbf{P(\text{Innocent} | \text{Jail})} \approx \mathbf{0.08} \quad \mathbf{!!}$

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

- ▶ More than 1% of Ste. Frequentiste goes to jail.
- ▶ On Isle Bayesienne, 10% are assumed guilty, but nobody goes to jail.
- ▶ The disagreement wasn't about math or how to interpret $P()$.
- ▶ What was it about?

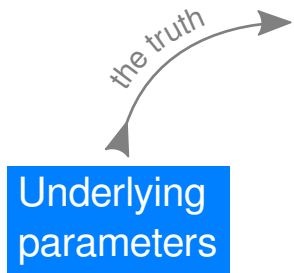
The islanders' concerns

- ▶ Frequentist cares about the rate of jailings among innocent people. Concern: **overall rate of false positive**
- ▶ Bayesian cares about the rate of innocence among jail inmates. Concern: **rate of error among positives**
- ▶ The Bayesian had to make an **assumption** about the overall probability of innocence.

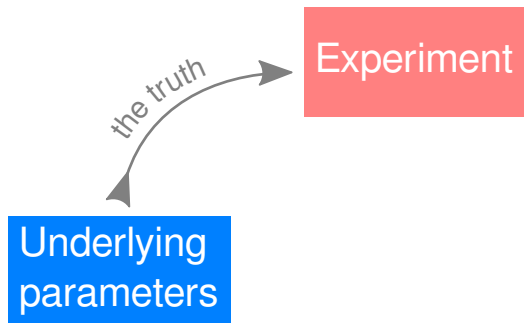
Quantifying uncertainty

Underlying
parameters

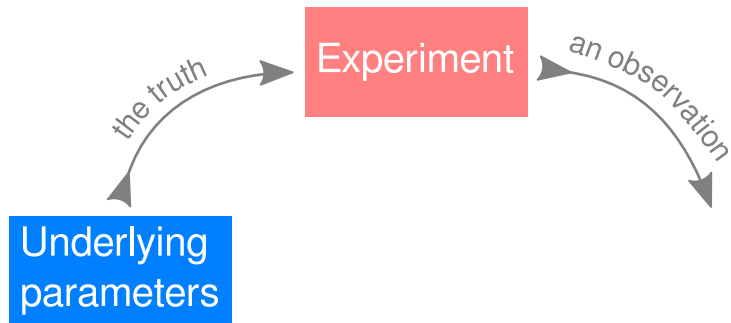
Quantifying uncertainty



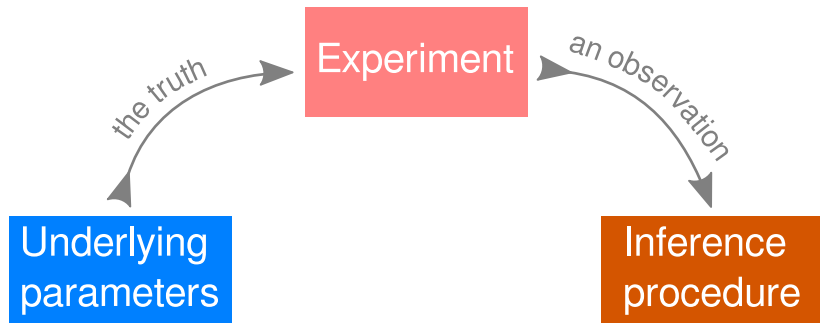
Quantifying uncertainty



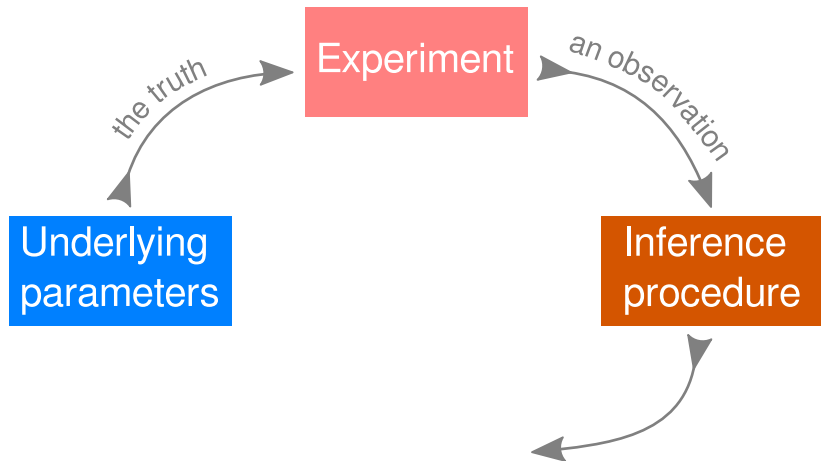
Quantifying uncertainty



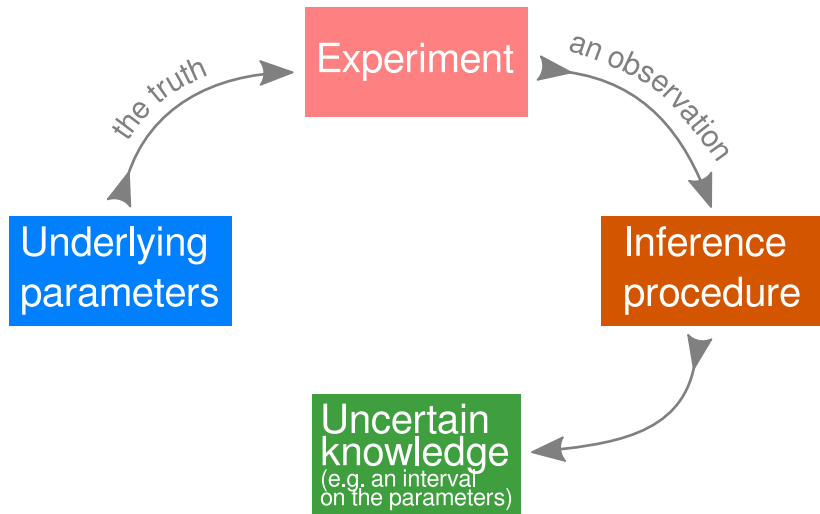
Quantifying uncertainty



Quantifying uncertainty



Quantifying uncertainty

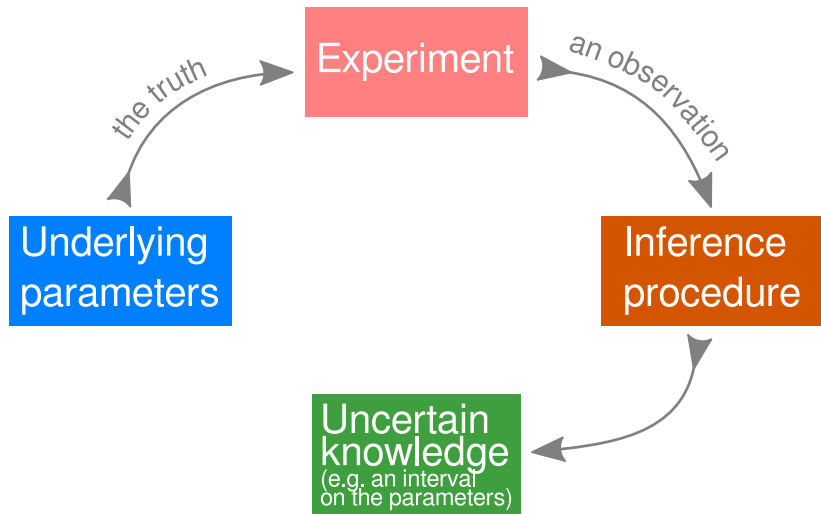


Jewel's Cookies

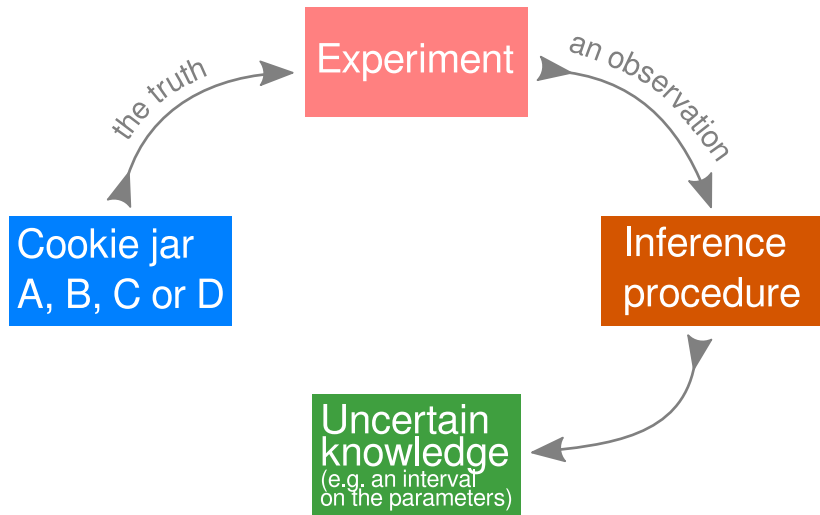
Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
total	100%	100%	100%	100%

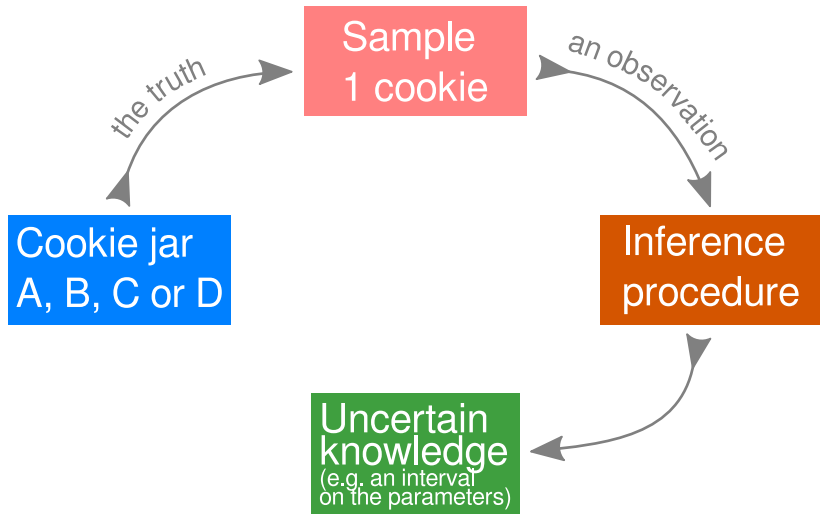
Quantifying cookie jar uncertainty



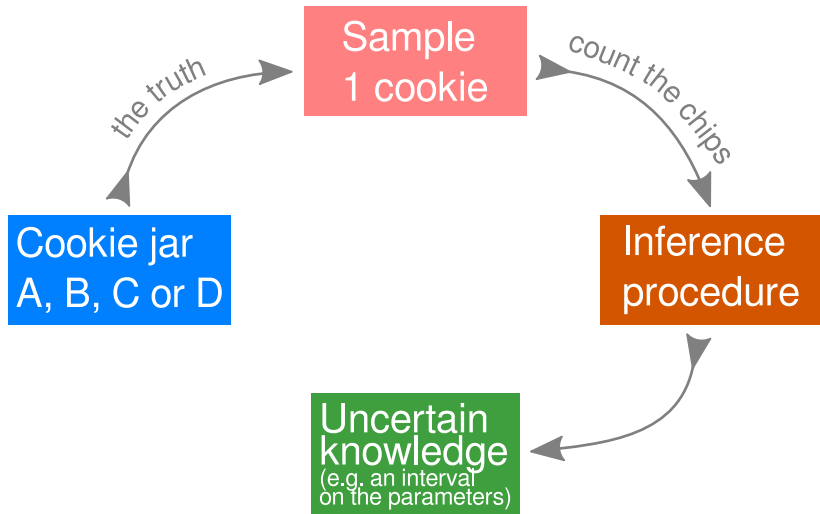
Quantifying cookie jar uncertainty



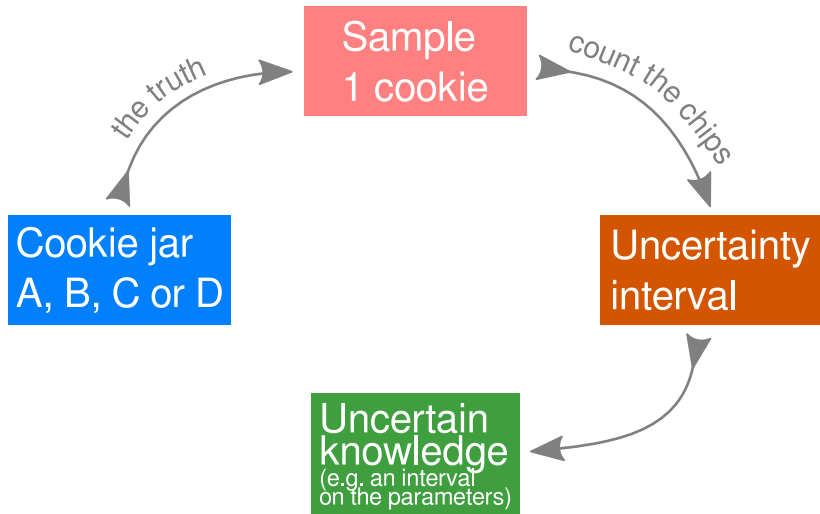
Quantifying cookie jar uncertainty



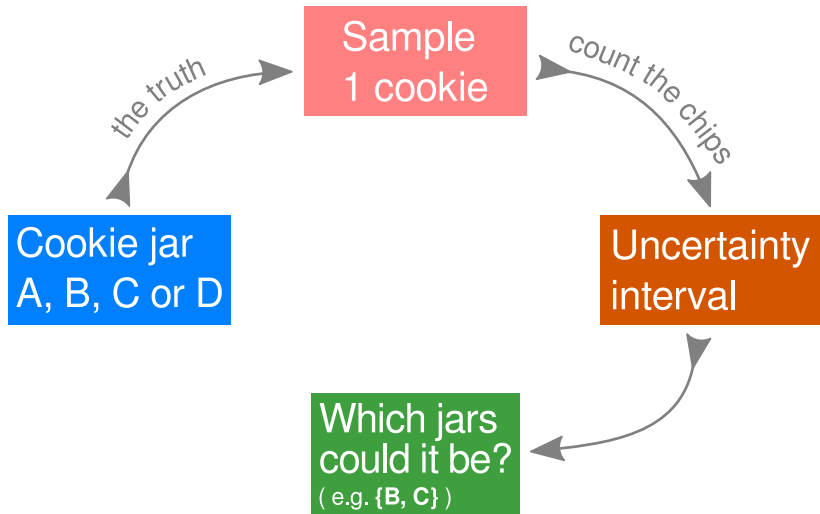
Quantifying cookie jar uncertainty



Quantifying cookie jar uncertainty



Quantifying cookie jar uncertainty



Frequentist inference

A 70% **confidence** interval method includes the correct jar with at least 70% probability **in the worst case, no matter what.**

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage				

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{ chips } \text{ jar })$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%			

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	25%		

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{ chips } \text{ jar })$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	49%		

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{ chips } \text{ jar })$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	69%		

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{ chips } \text{ jar })$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%		

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%	67%	

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{ chips } \text{ jar })$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%	87%	

Making 70% confidence intervals

Cookie jars **A**, **B**, **C**, **D** have 100 cookies each, but different numbers of chocolate chips per cookie:

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%	87%	70%

Bayesian inference

A 70% **credible** interval has at least 70% conditional probability of including the correct jar, **given the observation and the prior assumptions.**

Uniform prior

Our prior assumption: jars A, B, C, and D have equal probability.

Conditional probabilities

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
total	100%	100%	100%	100%

Conditional probabilities

$P(\text{chips} \mid \text{jar})$	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
total	100%	100%	100%	100%

Joint probabilities under uniform prior

$P(\text{chips} \cap \text{jar})$	A	B	C	D	$P(\text{chips})$
0	1/4	12/4	13/4	27/4	13.25%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%
total $P(\text{jar})$	25%	25%	25%	25%	100%

Joint probabilities under uniform prior

$P(\text{chips} \cap \text{jar})$	A	B	C	D	$P(\text{chips})$
0	1/4	12/4	13/4	27/4	13.25%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%
total $P(\text{jar})$	25%	25%	25%	25%	100%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1/4	12/4	13/4	27/4	13.25%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	1/4	19/4	20/4	70/4	27.50%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	70/4	24/4	0/4	1/4	23.75%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	28/4	20/4	0/4	1/4	12.25%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	57.1	40.8	0.0	2.0	100%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	<i>P(chips)</i>
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	57.1	40.8	0.0	2.0	100%
4	0/4	25/4	67/4	1/4	23.25%

Posterior probabilities under uniform prior

	A	B	C	D	$P(\text{chips})$
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	57.1	40.8	0.0	2.0	100%
4	0.0	26.9	72.0	1.1	100%

Posterior probabilities under uniform prior

$P(\text{jar} \mid \text{chips})$	A	B	C	D	$P(\text{chips})$
0	1.9	22.6	24.5	50.9	100%
1	0.9	17.3	18.2	63.6	100%
2	73.7	25.3	0.0	1.1	100%
3	57.1	40.8	0.0	2.0	100%
4	0.0	26.9	72.0	1.1	100%

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	
1	0.9	17.3	18.2	63.6	
2	73.7	25.3	0.0	1.1	
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	51%
1	0.9	17.3	18.2	63.6	
2	73.7	25.3	0.0	1.1	
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	
2	73.7	25.3	0.0	1.1	
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	64%
2	73.7	25.3	0.0	1.1	
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	82%
2	73.7	25.3	0.0	1.1	
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	82%
2	73.7	25.3	0.0	1.1	74%
3	57.1	40.8	0.0	2.0	
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	82%
2	73.7	25.3	0.0	1.1	74%
3	57.1	40.8	0.0	2.0	57%
4	0.0	26.9	72.0	1.1	

70% credible intervals

$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	82%
2	73.7	25.3	0.0	1.1	74%
3	57.1	40.8	0.0	2.0	98%
4	0.0	26.9	72.0	1.1	

70% credible intervals

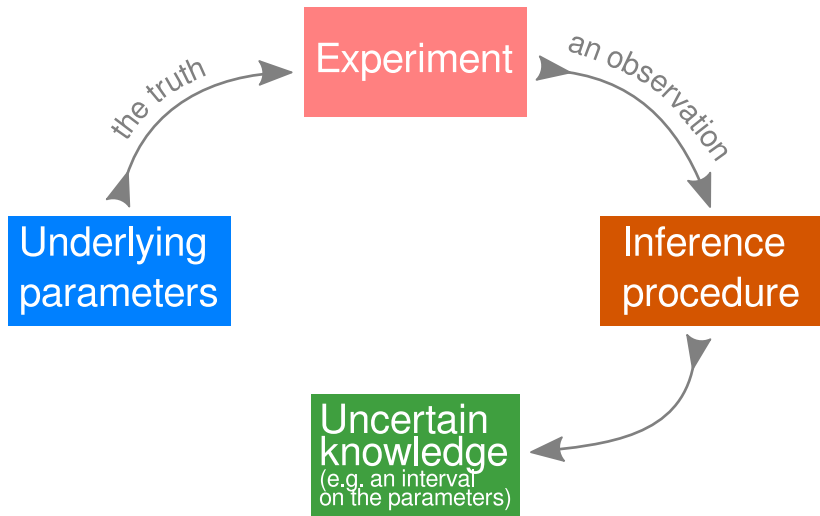
$P(\text{jar} \mid \text{chips})$	A	B	C	D	probability
0	1.9	22.6	24.5	50.9	75%
1	0.9	17.3	18.2	63.6	82%
2	73.7	25.3	0.0	1.1	74%
3	57.1	40.8	0.0	2.0	98%
4	0.0	26.9	72.0	1.1	72%

Confidence & credible intervals together

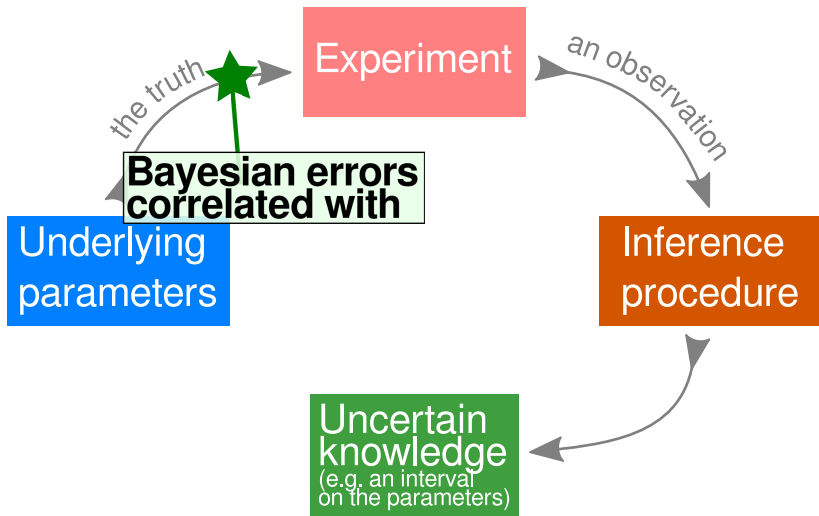
confidence	A	B	C	D	probability
0	1	12	13	27	0%
1	1	19	20	70	99%
2	70	24	0	1	99%
3	28	20	0	1	41%
4	0	25	67	1	99%
coverage	70%	88%	87%	70%	

credible	A	B	C	D	probability
0	1	12	13	27	75%
1	1	19	20	70	82%
2	70	24	0	1	74%
3	28	20	0	1	98%
4	0	25	67	1	72%
coverage	98%	20%	100%	97%	

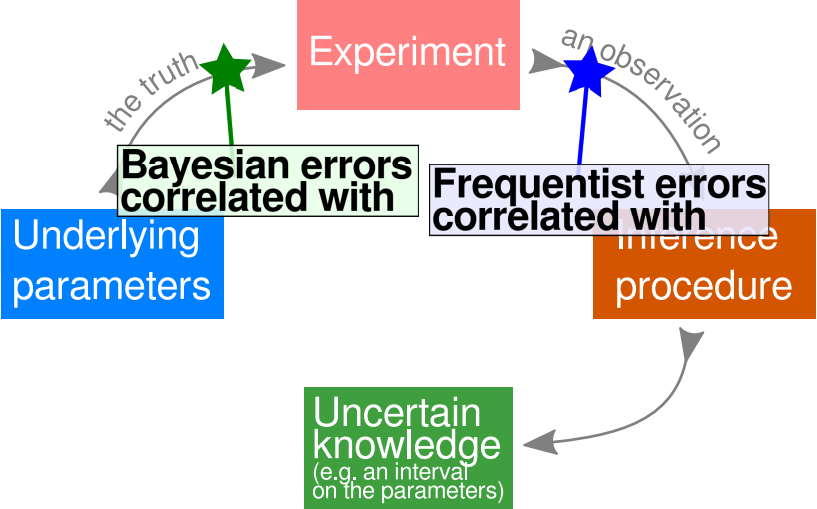
Correlation of error



Correlation of error



Correlation of error



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 is less likely to be true when the study

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 on field targets highly likely or searches for only on true relationships among and millions of hypothesis be postulated. Let us a for computational simulation circumscribed fields which is only one true relationship many that can be hypothesized. 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

ECONOMETRICA

VOLUME 47

NOVEMBER, 1979

NUMBER 6

THE IMPOSSIBILITY OF BAYESIAN GROUP DECISION MAKING WITH SEPARATE AGGREGATION OF BELIEFS AND VALUES

BY AANUND HYLLAND AND RICHARD ZECKHAUSER¹

Bayesian theory for rational individual decision making under uncertainty prescribes that the decision maker define independently a set of beliefs (probability assessments for the states of the world) and a system of values (utilities). The decision is then made by maximizing expected utility. We attempt to generalize the model to group decision making. It is assumed that the group's belief depends only on individual beliefs and the group's values only on individual values, that the belief aggregation procedure respects unanimity, and that the entire procedure guarantees Pareto optimality. We prove that only trivial (dictatorial) aggregation procedures for beliefs are possible.

1. INTRODUCTION

MANY DECISIONS MADE under uncertainty, indeed many important ones, are made by a group, be it a collection of friends, the Congress of the United States, or

Disagreement in the real world

- ▶ Avandia: world's #1 diabetes drug, approved in 1999.
- ▶ Sold by GlaxoSmithKline PLC
- ▶ Sales: \$3 billion in 2006
- ▶ In 2004, GSK releases results of many small studies.

GSK releases 42 small studies

Study	Avandia heart attacks	Control heart attacks
49632-020	2/391	1/207
49653-211	5/110	2/114
DREAM	15/2635	9/2634
49653-134	0/561	2/276
49653-331	0/706	0/325
⋮	⋮	⋮

In 2007, Dr. Nissen crashes the party

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

JUNE 14, 2007

VOL. 356 NO. 24

Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

ABSTRACT

BACKGROUND

Rosiglitazone is widely used to treat patients with type 2 diabetes mellitus, but its effect on cardiovascular morbidity and mortality has not been determined.

METHODS

We conducted searches of the published literature, the Web site of the Food and Drug Administration, and a clinical-trials registry maintained by the drug manufacturer (GlaxoSmithKline). Criteria for inclusion in our meta-analysis included a study duration of more than 24 weeks, the use of a randomized control group not receiving rosiglitazone, and the availability of outcome data for myocardial infarction and death from cardiovascular causes. Of 116 potentially relevant studies, 42 trials met the inclusion criteria. We tabulated all occurrences of myocardial infarction and death from cardiovascular causes.

RESULTS

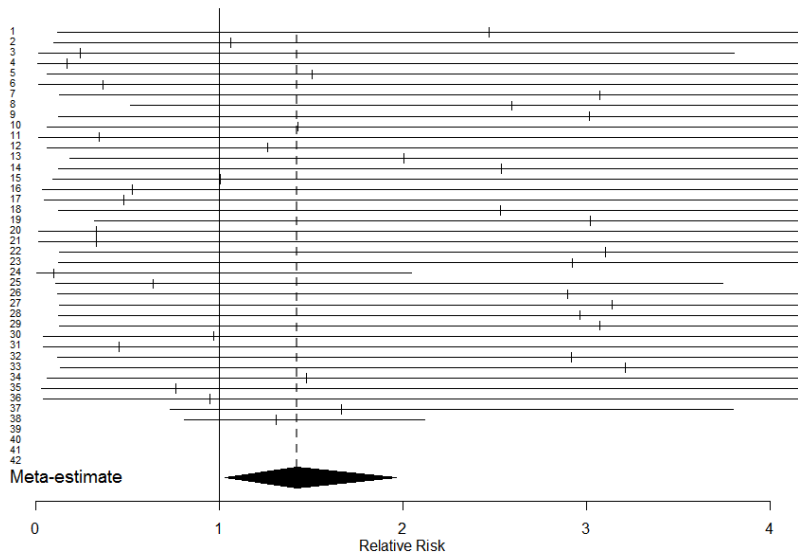
Data were combined by means of a fixed-effects model. In the 42 trials, the mean age of the subjects was approximately 56 years, and the mean baseline glycated hemoglobin level was approximately 8.2%. In the rosiglitazone group, as compared with the control group, the odds ratio for myocardial infarction was 1.43 (95% confidence interval [CI], 1.03 to 1.98; $P=0.03$), and the odds ratio for death from cardiovascular causes was 1.64 (95% CI, 0.98 to 2.74; $P=0.06$).

From the Cleveland Clinic, Cleveland. Address reprint requests to Dr. Nissen at the Department of Cardiovascular Medicine, Cleveland Clinic, 9500 Euclid Ave., Cleveland, OH 44195, or at nissens@ccf.org.

This article (10.1056/NEJMoa072761) was published at www.nejm.org on May 21, 2007.

N Engl J Med 2007;356:2457-71.
Copyright © 2007 Massachusetts Medical Society.

Frequentist inference



THE WALL STREET JOURNAL

DOJONES

TUESDAY, MAY 22, 2007 - VOL. CCXLIX NO. 119

**** \$1.00

DJIA 13542.88 ▼ 13.65 -0.1% NASDAQ 2578.79 ▲ 0.8% NIKKEI 17556.87 ▲ 0.9% DJSTOXX 50 3905.70 ▼ 0.3% 10-YR TREAS ▲ 4/32, yield 4.790% OIL \$66.27 ▲ \$1.33 GOLD \$662.90 ▲ \$1.90 EURO \$1.3470 YEN 121.45

What's News—

Business and Finance

World-Wide

U.S. employers are divided over the immigration bill, undermining its prospect of becoming law. Employers who rely on unskilled workers generally support the deal, but high-tech industries that need skilled workers complain that it doesn't give them the flexibility to recruit workers with the specific skills they need from abroad. **A1, A6**

Keckorian's Tracinda launched an overture for MGM Mirage's Bellagio Hotel and CityCenter project in Las Vegas, a volley that has put the whole company in play. **A3**

Glaxo shares slid after the New England Journal of Medicine released an analysis suggesting users of diabetes drug Avandia have a higher risk of heart attacks. **A1, D2**

EMI agreed to be bought by private-equity firm Terra Firma for \$4.73 billion, but the music company's shares rose 9.3% in a sign bidding may not be over. **A3**

Low's posted a 12% profit drop and cut its full-year outlook but said it will keep up an aggressive store-opening campaign. **A3**

Lebanon pounded a Palestinian camp in a second day of fighting. Artillery and tank fire engulfed a refugee camp outside Tripoli in violence that has killed at least 50 combatants and an unknown number of civilians. The Lebanese military surrounding the Nahr el-Balad camp sought to crush a militant al-Qaeda-inspired group holed up inside. Lebanon's worst internal violence since the 1975-1990 civil war rises fears tension could spread. In Beirut, a bomb rocked a shopping area in a mainly Sunni Muslim district, wounding at least four in the second explosion in two days. **A5**

The U.S. is bracing for a possible showdown with Russia and China over the establishment of a U.S. court to try suspects in the killing of Lebanon's ex-prime minister Hariri.

Iraq's military is drawing up plans to cope with any quick U.S. military pullout, the defense minister said, as an American official warned the Bush administration may reconsider its support if Iraqi leaders don't make major reforms by fall. Meanwhile, several mortar shells hit the Green Zone but caused no casualties.

U.S. troops raided safe houses south of Baghdad but failed to find three soldiers missing since May 12. A Florida doctor was convicted of

Two Jima Letters Of Young Japanese Are Home at Last

An American's Souvenir, They Had Sat on a Shelf, Solving a Family Mystery

By SEBASTIAN MOFFETT

KOBE, Japan—After the fighting died down in the Battle of Iwo Jima, Victor Voegelin, then 10 years old, was searching for the comrade when he saw a piece of thread poking out of the ground in a blown-out gun emplacement. The U.S. Navy petty officer pulled the thread, and found it was attached to a pack of letters, along with part of a ceramic sake cup and some cigarettes. He picked it all up and put everything in his bag.

Over the decades, Mr. Voegelin looked at the letters just three or four times. He couldn't read the Japanese script, and he always wanted to send them back to Japan. As he got older, "I started thinking about these letters," says Mr. Voegelin, "and thought that people around my age might be around who would want them."

Finally spurred by the release last year of the movie "Letters From Iwo Jima," he took action. He found that the letters had belonged to Tadashi Matsukawa, a Japanese sailor who was 23 when he died. Earlier this year, he sent them to Tadashi's brother, Masaji, at the same age as Mr. Voegelin.

MEDICAL DETECTIVE

Sequel for Vioxx Critic: Attack on Diabetes Pill

Glaxo Shares Plunge As Dr. Nissen Sees Risk To Heart From Avandia

By ANNA WILDE MATHESON

An analysis linking the widely used diabetes drug Avandia to a higher risk of heart attacks represents a serious blow to GlaxoSmithKline PLC and underscores how outside critics have been empowered to challenge big-selling drugs after the outcry over the withdrawn painkiller Vioxx.

Glaxo rang up more than \$3 billion in world-wide sales of Avandia last year. Its share price fell more than 7% after the New England Journal of Medicine released the analysis by prominent cardiologist Steven Nissen of the Cleveland Clinic, who helped raise early safety concerns about Vioxx. The analysis suggested that people on Avandia have a 43% higher chance of suffering a heart attack. Glaxo said it "strongly disagrees" with his conclusions, which come from

Drug in Demand Sales of GlaxoSmithKline's Avandia, in billions of pounds.



Note: £1 = \$1.97 at the current rate; includes sales of Avandamet and Avandaryl. Source: the company

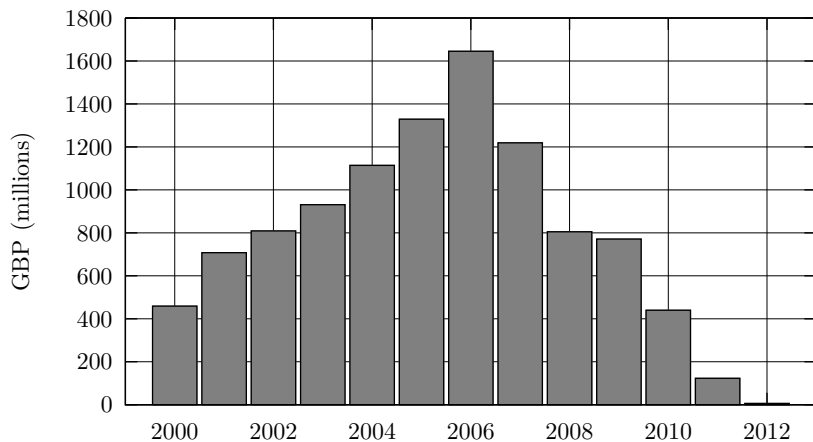
and Drug Administration should have acted faster to alert the public about possible risk from Avandia. Glaxo performed its own meta-analysis, which also showed a potential danger. It shared an early version of it with the FDA in September 2005 and a more complete one in August 2006. The findings weren't reflected on the U.S. label, which is supposed to give a comprehensive review of the drug's risks. Robert Meyer, head of the FDA office that oversees diabetes drugs, said the agency is still working on its analysis. "We have other data that suggests we



Steven Nissen

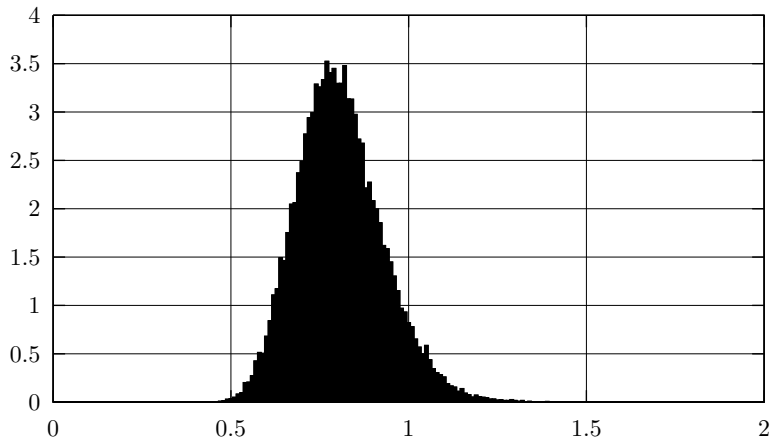
GlaxoSmithKline loses \$12 billion

Avandia worldwide sales



Bayesian inference disagrees, for risk **ratio**

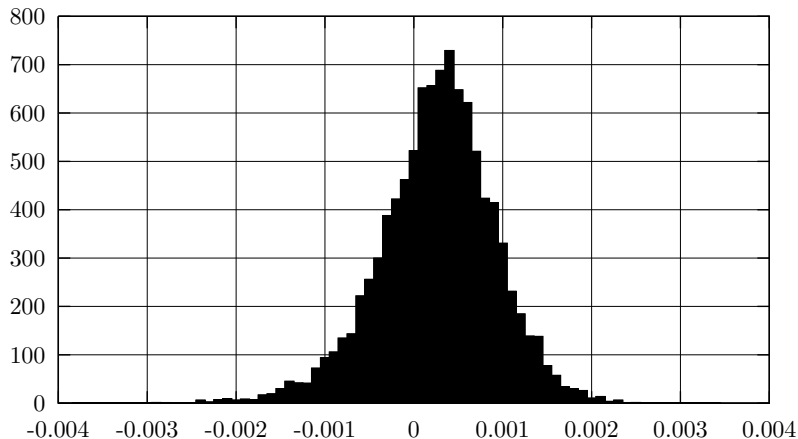
P.D.F. on Avandia's risk ratio for heart attack



(Joint work with Joshua Mandel, Children's Hospital Informatics Program)

Or does it? Here, risk **difference** model

P.D.F. on Avandia's risk difference for heart attack

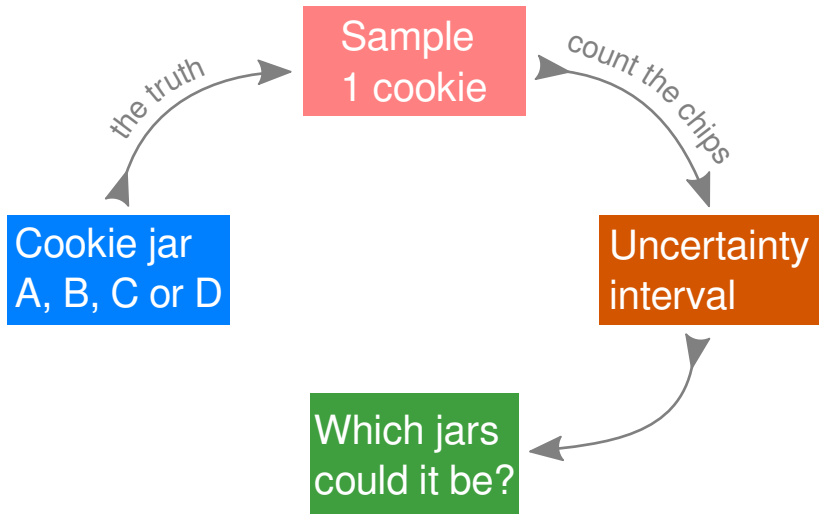


(Joint work with Joshua Mandel, Children's Hospital Informatics Program)

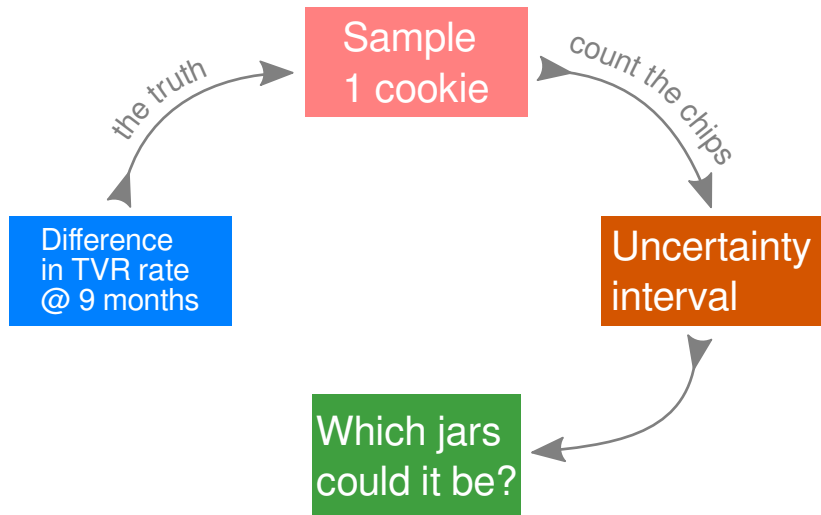
The TAXUS ATLAS Experiment

- ▶ Boston Scientific proposed to show that new heart stent was not “inferior” to old heart stent, with 95% confidence.
- ▶ Inferior means three percentage points more “bad” events.
 - ▶ CONTROL 7% vs. TREATMENT 10.5% \Rightarrow inferior
 - ▶ CONTROL 7% vs. TREATMENT 9.5% \Rightarrow non-inferior.

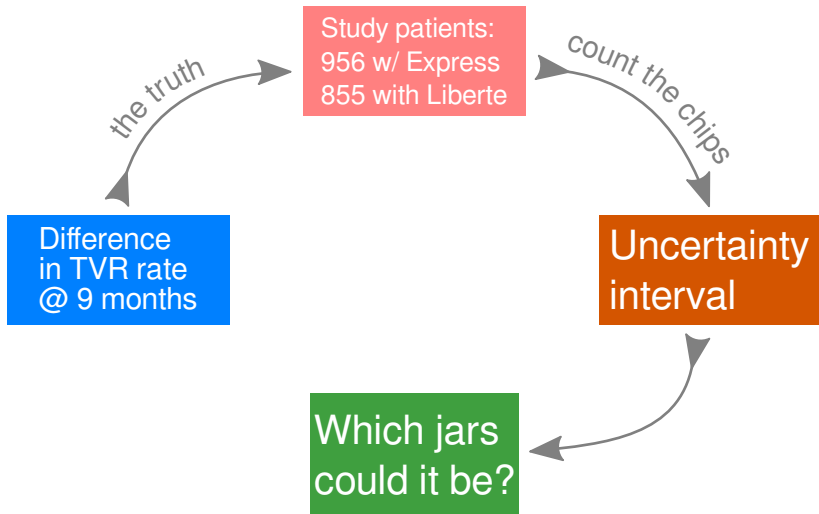
Measuring non-inferiority of coronary stents



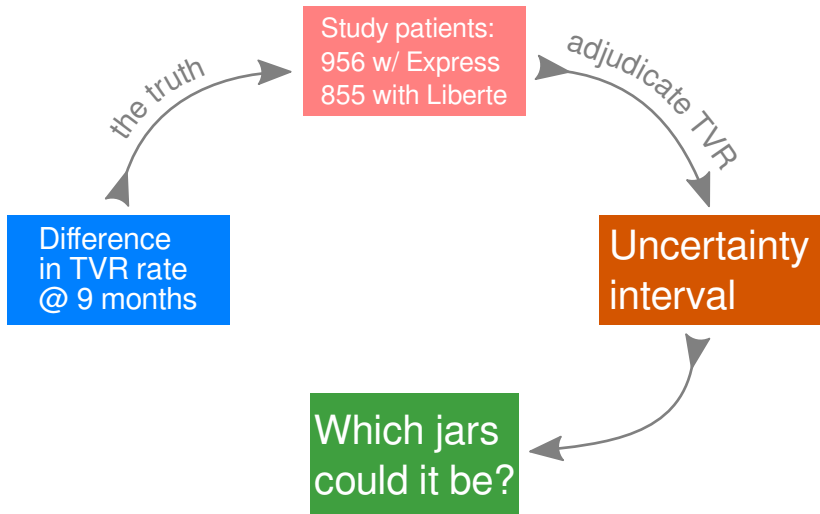
Measuring non-inferiority of coronary stents



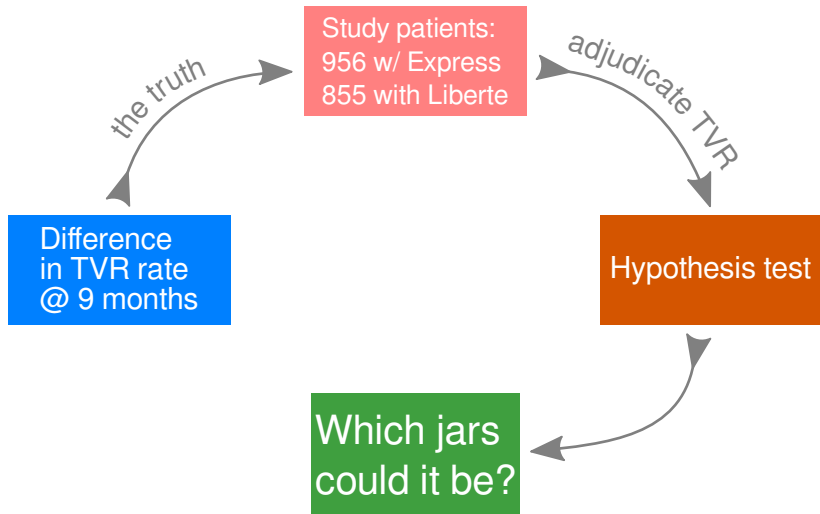
Measuring non-inferiority of coronary stents



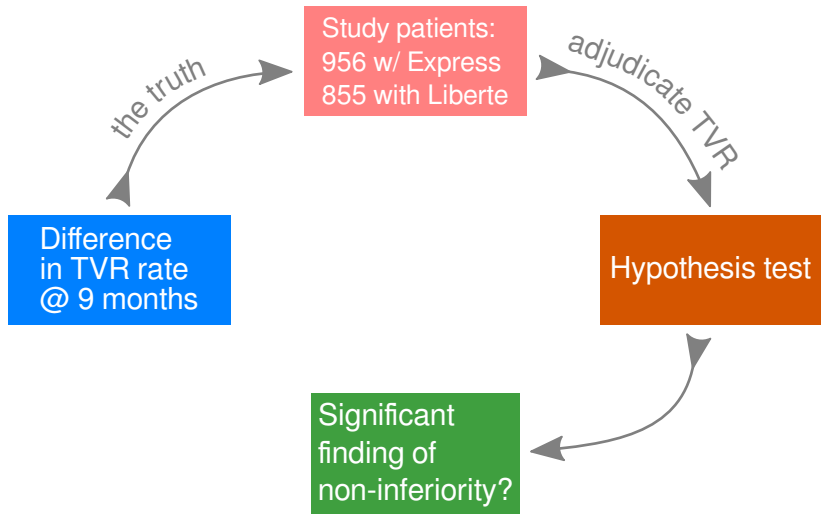
Measuring non-inferiority of coronary stents



Measuring non-inferiority of coronary stents



Measuring non-inferiority of coronary stents



ATLAS Results (May 2006)

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 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. Student *t* test was used to compare independent continuous variables, while chi-square or Fisher exact test was used to compare proportions.

p-value

$p < 0.05 \rightarrow 95\%$ confidence interval excludes inferiority

The problem

	Event	No event	Total
Control	67	889	956
Treatment	68	787	855
Total	135	1676	1811

The problem

	Event	No event	Total
Control	67	889	956
Treatment	68	787	855
Total	135	1676	1811

- ▶ With uniform prior on rates,
 $Pr(\textit{inferior} | \textit{data}) \approx 0.050737979\dots$
- ▶ **Posterior probability of non-inferiority is less than 95%.**

ATLAS trial solution

- ▶ Confidence interval: approximate *each binomial separately* with a normal distribution. Known as Wald interval.

- ▶ $p = \int_{0.03}^{\infty} \mathcal{N}\left(\frac{i}{m} - \frac{j}{n}, \frac{i(m-i)}{m^3} + \frac{j(n-j)}{n^3}\right) \approx 0.0487395 \dots$

- ▶ $p < 0.05 \rightarrow$ **success**

The ultimate close call

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

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
	65	66	67	68	69

TVR (Liberte)

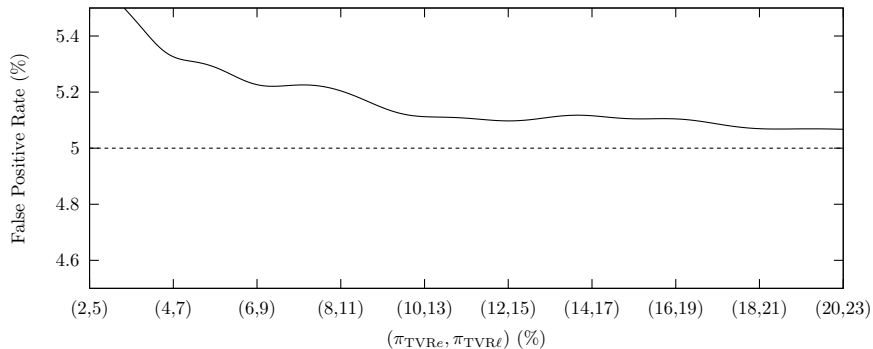
TVR (Express)

confidence	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%	87%	70%

confidence	A	B	C	D
0	1	12	13	27
1	1	19	20	70
2	70	24	0	1
3	28	20	0	1
4	0	25	67	1
coverage	70%	88%	87%	70%
false positive rate	30%	12%	13%	30%

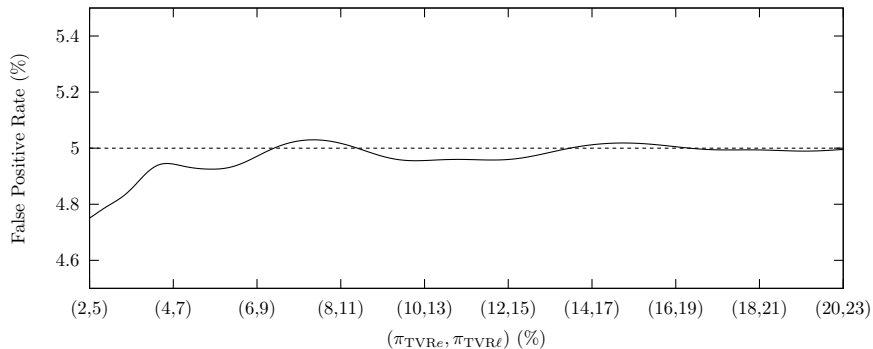
The Wald interval undercovers

False Positive Rate of ATLAS non-inferiority test along critical line



Better approximation: score interval

False Positive Rate of maximum-likelihood z -test along critical line



Other methods all yield failure

Method	p -value or confidence bound	Result
Wald interval	$p = 0.04874$	Pass
z-test, constrained max likelihood standard error	$p = 0.05151$	Fail

Other methods all yield failure

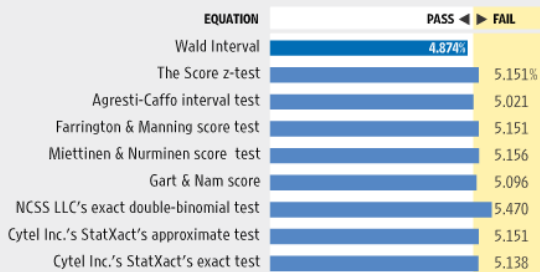
Method	<i>p</i> -value or confidence bound	Result
Wald interval	$p = 0.04874$	Pass
z-test, constrained max likelihood standard error	$p = 0.05151$	Fail
z-test with Yates continuity correction	$c = 0.03095$	Fail
Agresti-Caffo I_4 interval	$p = 0.05021$	Fail
Wilson score	$c = 0.03015$	Fail
Wilson score 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?

Degree of Certainty

Medical studies define success or failure in testing a hypothesis by calculating a degree of certainty, known as the p-value. The p-value must be less than 5% for the results to be considered significant. Boston Scientific's study, which used a statistical method called a Wald Interval, produced a p-value below 5%. But using 16 other methods turned up a p-value greater than 5%. Here are some of the p-values that resulted from the data in the study, using those different methodologies.

Source: WSJ research



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 worldwide.

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



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

by statisticians for exaggerating the certainty of research results. Rivals Medtronic Inc. and Abbott Laboratories have used the same equation in stent studies.

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

The statistician says. . .

- ▶ “ $np > 5$, therefore, the Central Limit Theorem applies and a Gaussian approximation is appropriate.”
- ▶ “We had even more data points than we powered the study for, so there was adequate safety margin.”
- ▶ “‘Exact’ tests are too conservative.”

StatXact calculates an “exact” test

“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. . .”

StatXact calculates an “exact” test

“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. . .”

Graphing the coverage

- ▶ **Problem:** hard to calculate a million “exact” p-values
- ▶ (StatXact: about 10 minutes each)
- ▶ **Contribution:** method for calculating whole tableau
- ▶ Calculates all p-values in time for “hardest” square
- ▶ **Trick:** calculate in the right order, cache partial results

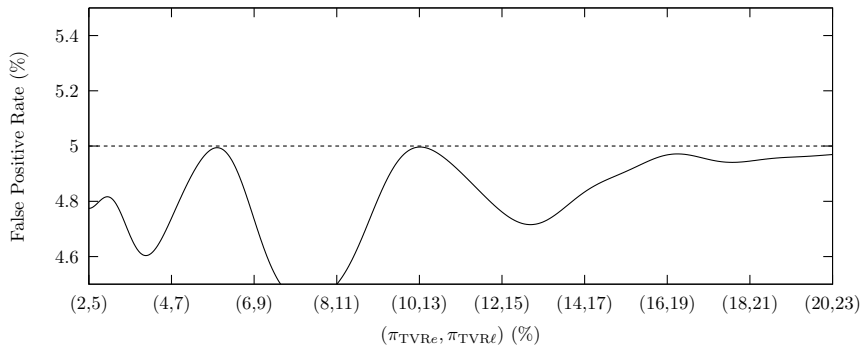
StatXact calculates an “exact” test

“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. . .”

StatXact calculates an “exact” test

“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. . . .”

Type I rate of StatXact 8 non-inferiority test (Berger Boos-adjusted Chan)



Summing up

- ▶ Bayesian and frequentist schools have much in common.
- ▶ This “battle” has been going on a long time.

An abbreviated history

- ▶ **1760**: Daniel Bernoulli shows smallpox vaccination is a cause of immunity, using “inverse probability.”
- ▶ **1763**: Bayes’s work presented posthumously
- ▶ **1800s**: Much “inverse” work. Struggles to formalize.
- ▶ **1933**: Kolmogorov axioms
- ▶ **1930s–60s**: Dramatic spread of “frequentist” techniques
- ▶ **1950s–**: Return of inverse (now “Bayesian”) techniques

20. One of the most ancient problems in probability is concerned with the gradual diminution of the probability of a past event, as the length of the tradition increases by which it is established. Perhaps the most famous solution of it is that propounded by Craig in his *Theologiae Christianae Principia Mathematica*, published in 1699.

“Craig,” says Todhunter, “concluded that faith in the Gospel so far as it depended on oral tradition expired about the year 880, and that so far as it depended on written tradition it would expire in the year 3150. Peterson by adopting a different law of diminution concluded that faith would expire in 1789.”²

²In the *Budget of Paradoxes* De Morgan quotes Lee, the Cambridge Orientalist, to the effect that Mahometan writers, in reply to the argument that the Koran has not the evidence derived from Christian miracles, contend that, as evidence of Christian miracles is daily weaker, a time must at last arrive when it will fail of affording assurance that they were miracles at all: whence the necessity of another prophet and other miracles.

Final thoughts

- ▶ Bayesian and frequentist are families of techniques. . .
... not tribes of people.
- ▶ **What's important:** say what we're trying to infer, how we get there, what we care about.
- ▶ Reasoning under uncertainty means compromising.
- ▶ Questions? keithw@mit.edu