Basic Bayes: I

Steven Goodman, MD, PhD

Johns Hopkins University

Baltimore, USA

sgoodman@jhmi.edu

A short research quiz

A well done study is reported on a new electrical stimulator for pain control, and the authors state that it has turned out, somewhat surprisingly (i.e.they thought this would have no more than a 25% chance of being true before the experiment), to be effective in reducing migraine pain, risk □=15%, 95% CI: 0 to 30%, *p*=0.05. The probability that this association is real is:

a.) < 75%

b.) 75% to 94.99...%

c.) ≥ 95%

Things I won't say

- That if we turn to Bayesian methods, all our problems will go away.
- That the only "right thinkers" in the statistics world are Bayesian.
- That the Bayesian approach doesn't have difficulties.

Things I will say

- That if we turn to Bayesian methods, difficult issues will be discussed in the right way by the right people.
- Some of the dilemmas that FDA decision-makers face are artifacts of the statistical methods they use, and not due to demands of the scientific method.
- That the Bayesian perspective provides the best way to think about evidence.

Things identified as cancer risks

(Altman and Simon, JNCI, 1992)

- Electric Razors
- Broken Arms (in women)
- Fluorescent lights
- Allergies
- Breeding reindeer
- Being a waiter
- Owning a pet bird
- Being short
- Being tall
- Hot dogs

Having a refrigerator!!

S. Goodman, FDA Bayes Workshop

A16 THE NEW YORK TIMES NATIONAL WEDNESDAY, JANUARY 6, 1999

Magnets Lessen Foot Pain Of Diabetics, a Study Finds

By HOLCOMB B. NOBLE

In one of the first scientific studies of the centuries-old and highly debated use of magnets for treatment of medical disorders, a New York neurologist reported today that he had significantly lessened the foot pain that afflicts millions of diabetics.

Dr. Michael I. Weintraub, a clinical professor of neurology at New York Medical College, emphasized that his study was small, involving only 24 patients, and must be regarded as preliminary to much more clinical research. But he said that the early results were clear and that the treatment ought to be put to use immediately, provided the correct magnets are used and the treatment is limited to the types of pain that have been studied.

The study, which appears in this ... the American Jour-

A finding that runs counter to many previous studies.

fects have established that they actually work.

In November 1997, reporting in the Archives of Physical and Rehabilitation Medicine, Dr. Carlos Vallbona of the Baylor College of Medicine in Houston said that he applied low-intensity magnets to his own knee pain and that the pain was gone ... minutes. He then did a small study ... patients with post-polio-syndrome pain. One group was exposed to small magnets, the other to sham magnets. The patients with real magnets reported a 50 percent reduction in pain, while the others reported less than 10 percent.

... hands ... other ... and often ... hands and feet.

In July 1997, Dr. ... a four-month study of ... diabetic and nondiabetic ... Twenty-four patients with ... foot pain caused by diabetes, multiple myeloma, uremia, ischemia, ... pus and alcoholism were enrolled in a randomized placebo study.

In the first month, each patient ...

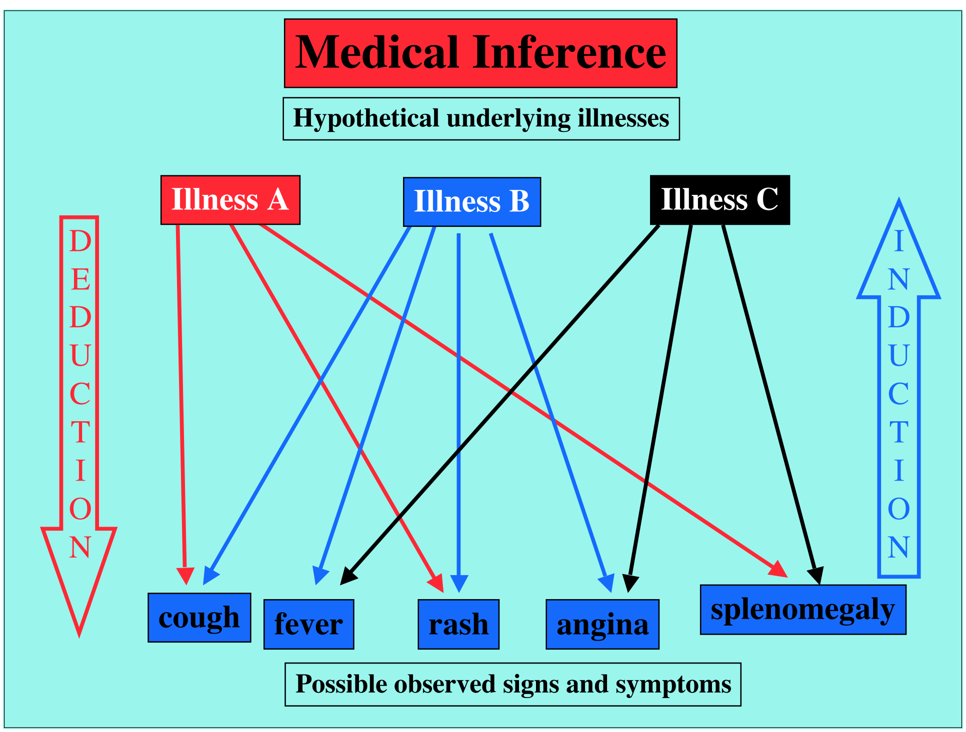
"We have no idea how or why the magnets work."

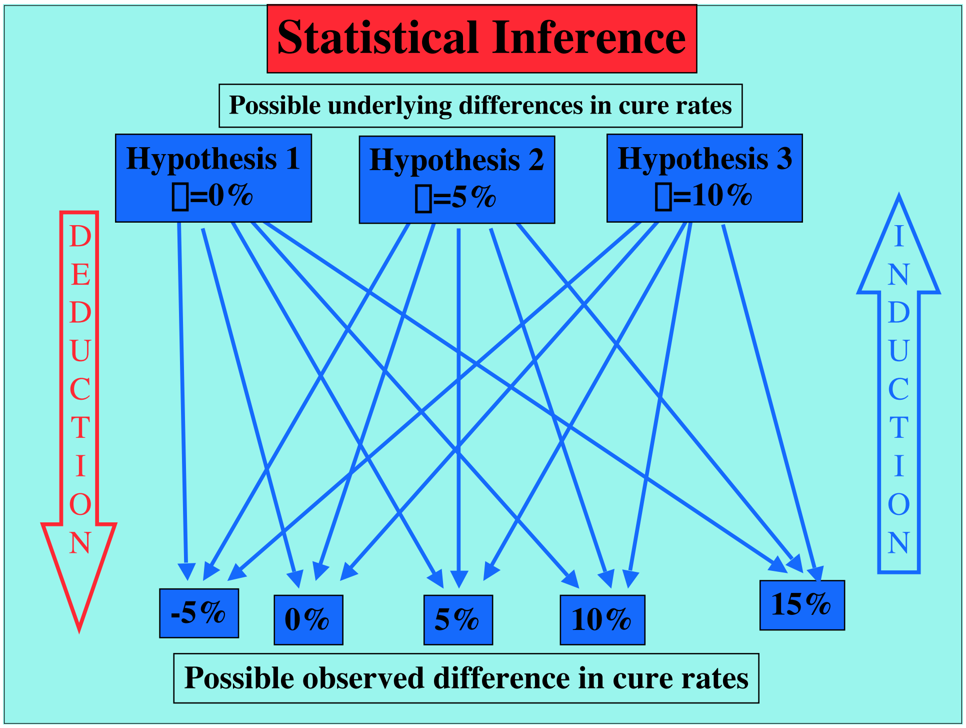
"A real breakthrough…"

"…the [study] must be regarded as preliminary…."

"But…the early results were clear and… the treatment ought to be put to use immediately."


Inference


Medical Inference
Hypothetical underlying illnesses
Illness A
Illness B
Illness C
DEDUCTION
INDUCTION
cough
fever
rash
angina
splenomegaly
Possible observed signs and symptoms



Statistical inference

- There is only one formal, coherent calculus of statistical inference: Bayes Theorem.
- “Traditional” statistical rules of inference are a collection of principles and conventions to avoid errors over the long run. *They do not tell us how likely our claims are to be true, nor do they easily apply to individual results.*

S. Goodman, FDA Bayes Workshop

Bayes Theorem

Odds that hypothesis is true before seeing the data (Prior Odds) ↔ "Subjective" component

X

Bayes factor ↔ Data component (Evidence)

Final (posterior) odds that the hypothesis is true)

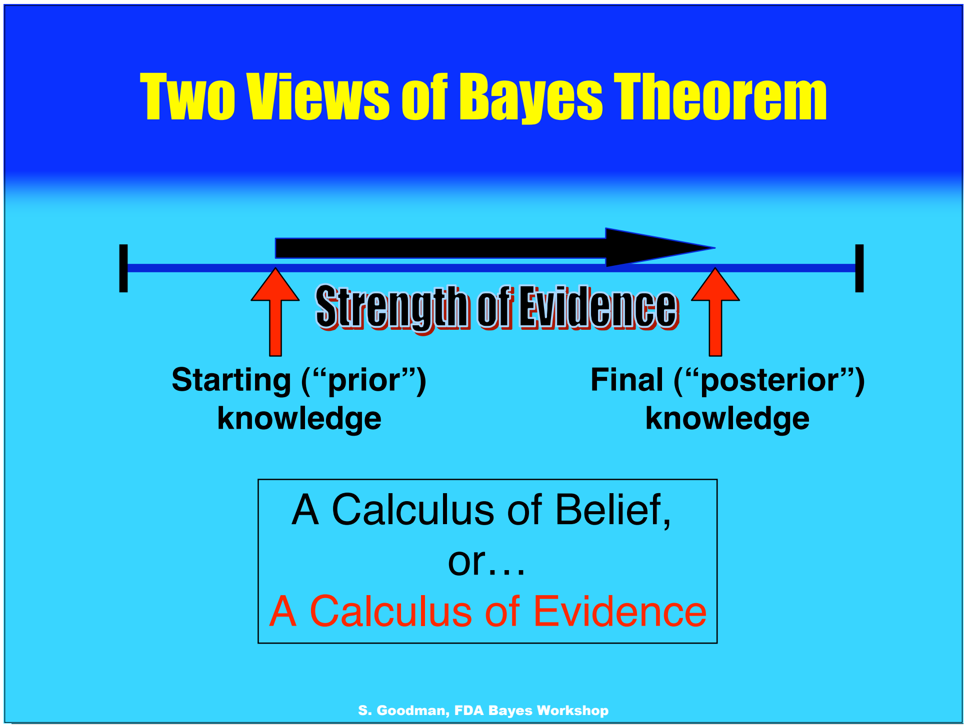
S. Goodman, FDA Bayes Workshop

Bayes Theorem

$$\underbrace{\frac{\Pr(H_0 \mid \text{Data})}{\Pr(H_1 \mid \text{Data})}}_{\textit{Post–test Odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\textit{Pre–test Odds}} \times \underbrace{\frac{\Pr(\text{Data} \mid H_0)}{\Pr(\text{Data} \mid H_1)}}_{\textit{Likelihood Ratio}}$$

a.k.a. Bayes factor

S. Goodman, FDA Bayes Workshop

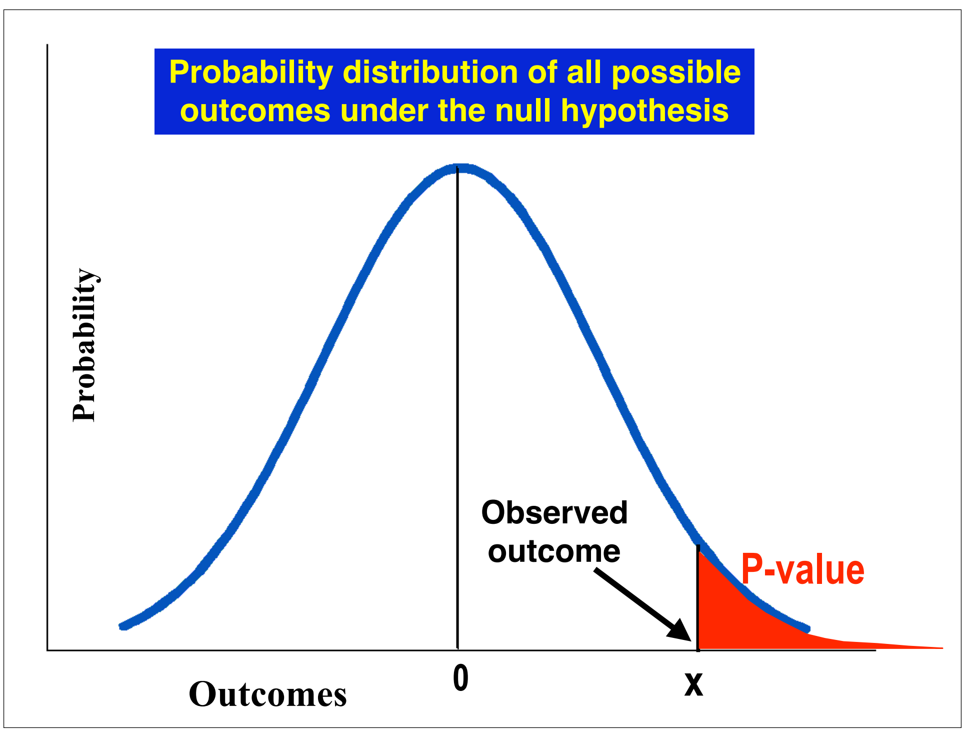

Two Views of Bayes Theorem
Strength of Evidence
Starting ("prior") knowledge
Final ("posterior") knowledge
A Calculus of Belief,
or…
A Calculus of Evidence
S. Goodman, FDA Bayes Workshop


P-values

RA Fisher on statistical education

"I am quite sure it is only personal contact with ... the natural sciences that is capable to keep straight the thought of mathematically-minded people...I think it is worse in this country [the USA] than in most, though I may be wrong. Certainly there is grave confusion of thought. We are quite in danger of sending highly trained and intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort." 1958

S. Goodman, FDA Bayes Workshop



Meaning of the p-value

Probability?

Plausibility?

Possibility?

Publish!!!

S. Goodman, FDA Bayes Workshop

The P-value is...

- The probability of getting a result as or more extreme than the observed result, if the null hypothesis (of chance) were true.
- Since the p-value is calculated *assuming the null hypothesis to be true*, it cannot represent the *probability of the truth of the null hypothesis.*

S. Goodman, FDA Bayes Workshop

The P-value is not....

- "The probability of the null hypothesis."
- "The probability that you will make a Type I error if you reject the null hypothesis."
- "The probability that the observed data occurred by chance."
- "The probability of the observed data under the null hypothesis."

S. Goodman, FDA Bayes Workshop

FDA Discussion

(Fisher, CCT, 20:16-39,1999)

L. Moyé, MD, PhD

"What we have to wrestle with is how to interpret p-values for secondary endpoints in a trial which frankly was negative for the primary. …In a trial with a positive endpoint…you haven't spent all of the alpha on that primary endpoint, and so you have some alpha to spend on secondary endpoints….In a trial with a negative finding for the primary endpoint, you have no more alpha to spend for the secondary endpoints."

S. Goodman, FDA Bayes Workshop

FDA Discussion, cont.

(Fisher, CCT, 20:16-39,1999)

Dr. Lipicky: What are the p-values needed for the secondary endpoints? …Certainly we're not talking 0.05 anymore. …You're out of this 0.05 stuff and I would have like to have seen what you thought was significant and at what level…

What p-value tells you that it's there study after study?

Dr. Konstam: …what kind of statistical correction would you have to do that survival data given the fact that it's not a specified endpoint? I have no idea how to do that from a mathematical viewpoint.

S. Goodman, FDA Bayes Workshop

Likelihood

Definition of Likelihood

- The degree to which a hypothesis predicts the data (probability) is proportional to the support that the data gives the hypothesis (likelihood).

If **Pr(Data | Hypothesis) = p**

Then

Likelihood (Hypothesis | Data) = c x **p**

where c = arbitrary constant

Bayes Theorem

$$\underbrace{\frac{\Pr(H_0 \mid \text{Data})}{\Pr(H_1 \mid \text{Data})}}_{\textit{Post–test Odds}} = \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\textit{Pre–test Odds}} \times \underbrace{\frac{\Pr(\text{Data} \mid H_0)}{\Pr(\text{Data} \mid H_1)}}_{\textit{Likelihood Ratio}}$$

$$\textit{Likelihood Ratio} = \overbrace{\frac{L(H_0 \mid \text{Data})}{L(H_1 \mid \text{Data})}}$$

Bayes factor vs. P-value

P-value	Bayes factor
Non-comparative	Comparative
Observed + hypothetical data	Only observed data
Alternative hypothesis implicit, partly data-defined	Alternative hypothesis explicit, pre-defined
Evidence only negative	Evidence negative or positive
Sensitive to stopping rules	Insensitive to stopping rules
No formal justification or interpretation	Formal justification and interpretation

Calibrating LRs

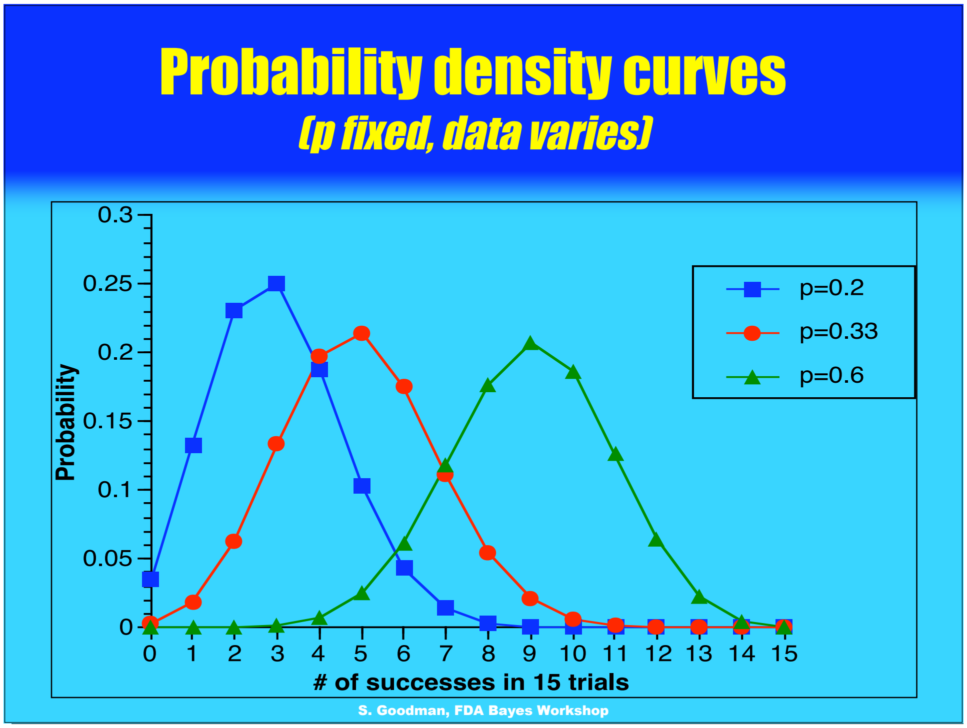
Strength of Evidence	LR	Final probability when prior probability = 25%	50%	75%
Zero	1	25	50	75
Moderate	5	62	83	94
Mod/Strong	10	77	91	97
Strong	20	83	95	98
Very Strong	40	93	97.5	99
Very Strong	80	96	99	99.6

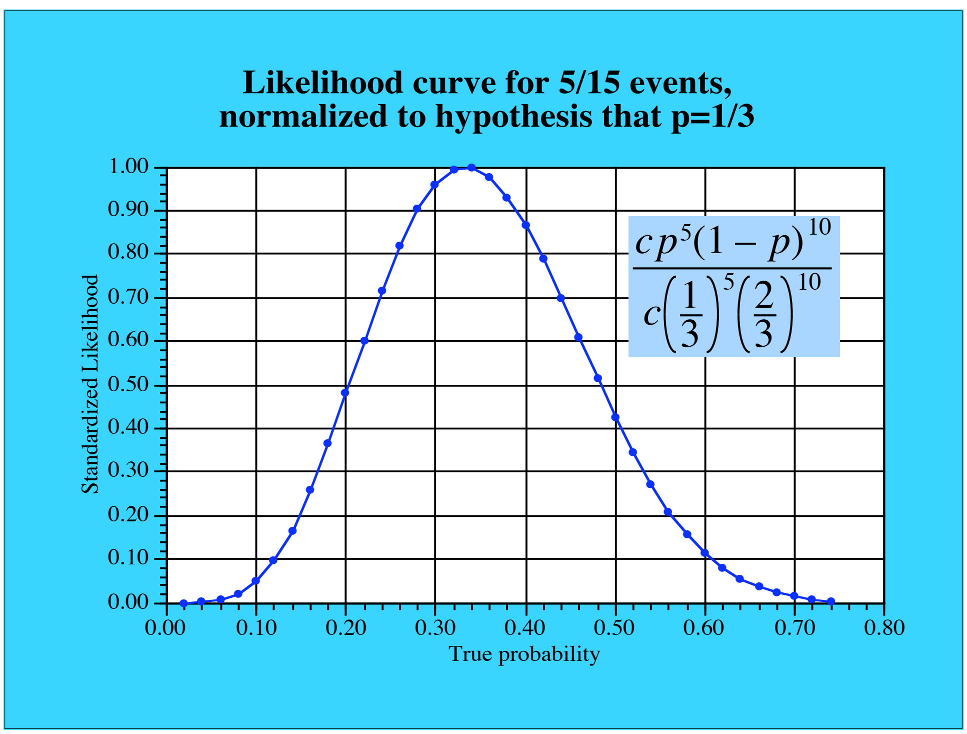
Examples of hypotheses

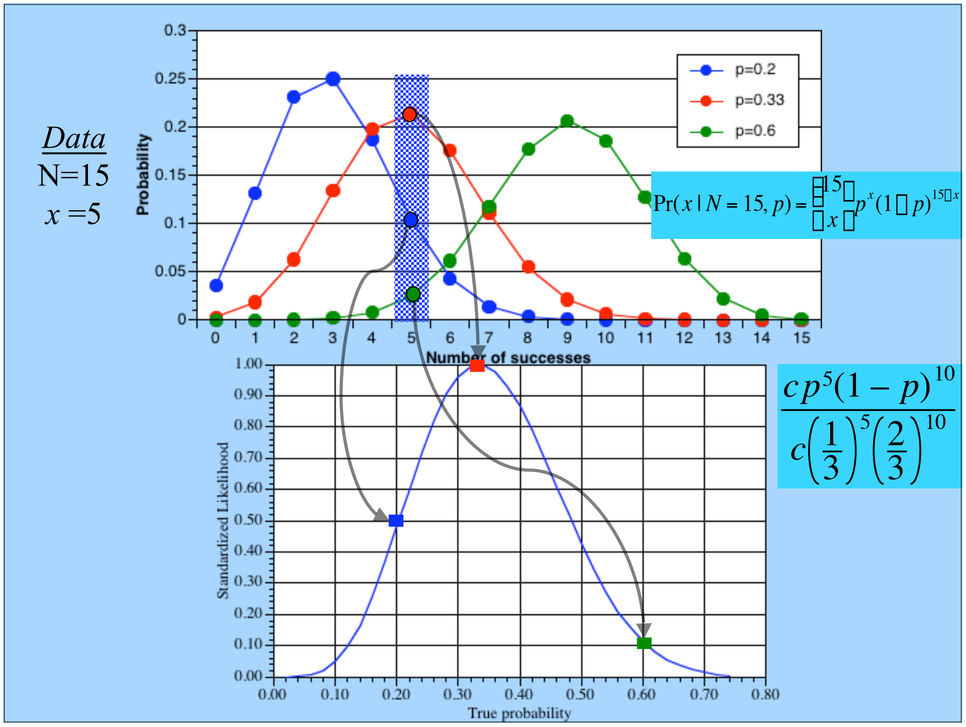
- Cure rate = 15% (Simple)
- Cure rate > 15% (Composite)
- Treatment difference = 0 (Simple)
- Treatment is beneficial (Composite)
- Treatment is harmful (Composite)

S. Goodman, FDA Bayes Workshop

Understanding Likelihood Functions

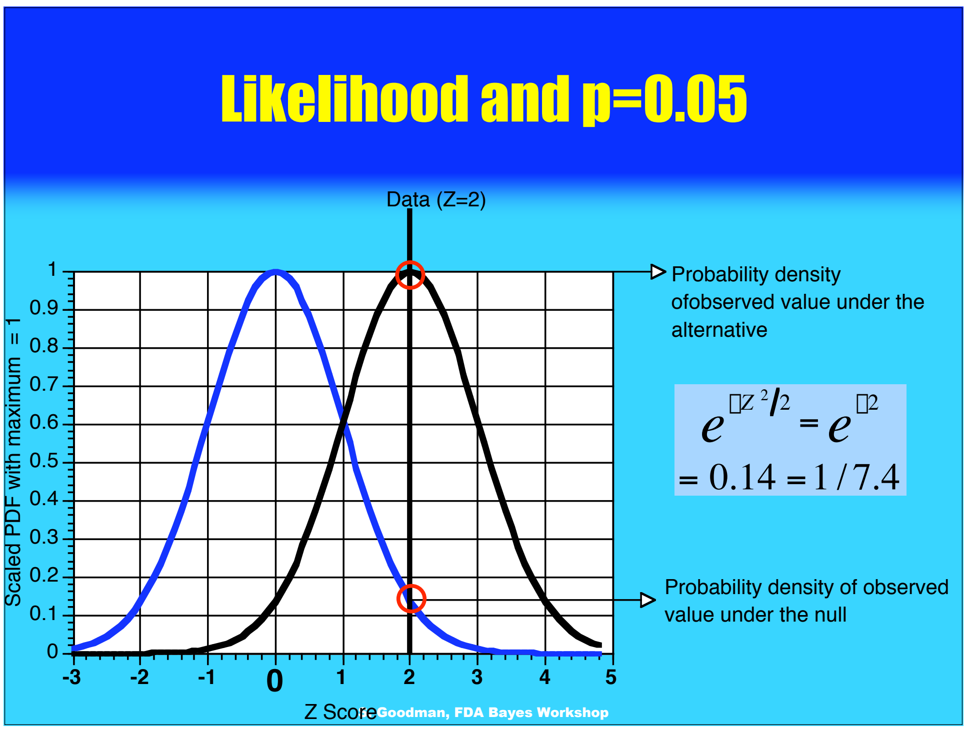

Probability density curves
(p fixed, data varies)
p=0.2
p=0.33
p=0.6
Probability
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
# of successes in 15 trials
S. Goodman, FDA Bayes Workshop


Likelihood curve for 5/15 events,
normalized to hypothesis that p=1/3
$$\frac{c\,p^5(1-p)^{10}}{c\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^{10}}$$
Standardized Likelihood
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80
True probability



$$\Pr(x \mid N=15, p) = \binom{15}{x} p^x (1-p)^{15-x}$$

$$\frac{c\,p^5(1-p)^{10}}{c\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^{10}}$$



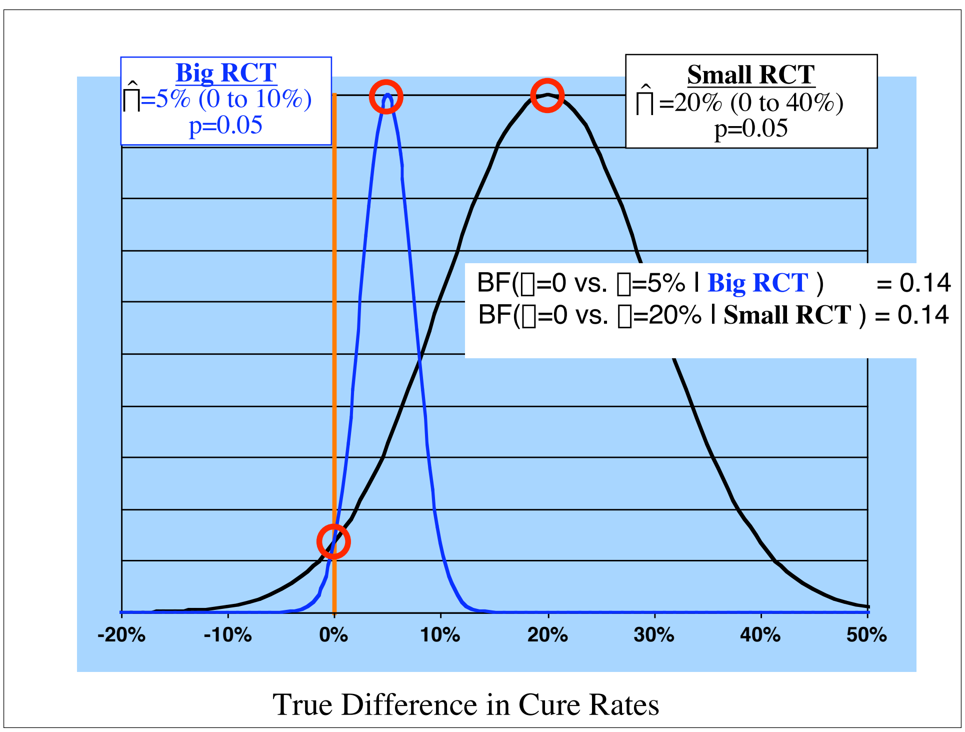
$$e^{-Z^2/2} = e^{-2} = 0.14 = 1/7.4$$

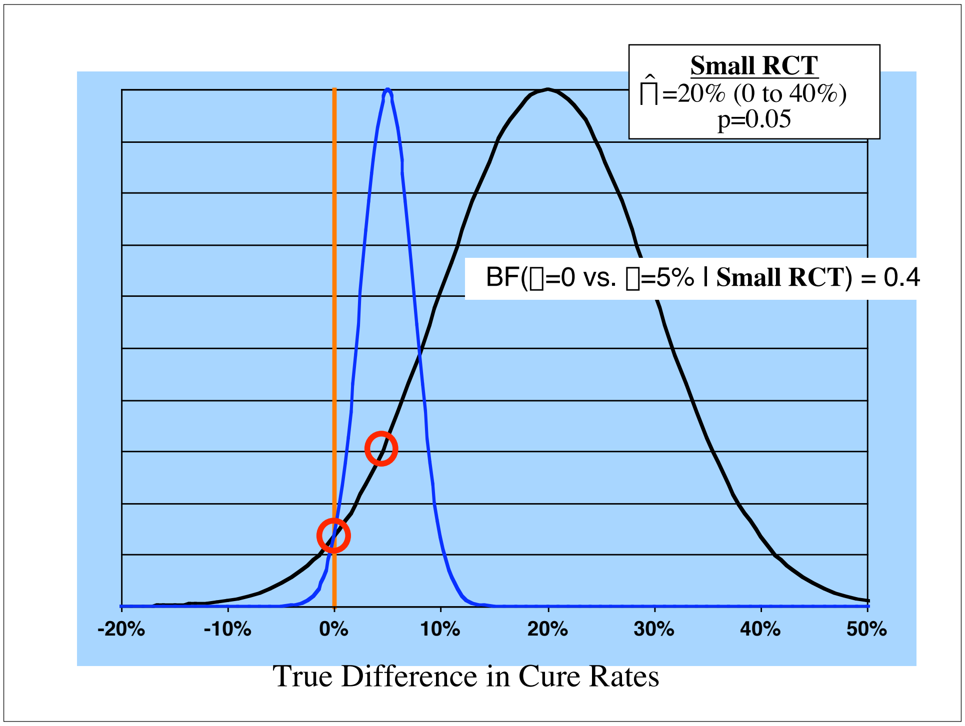
Standardized gaussian likelihood

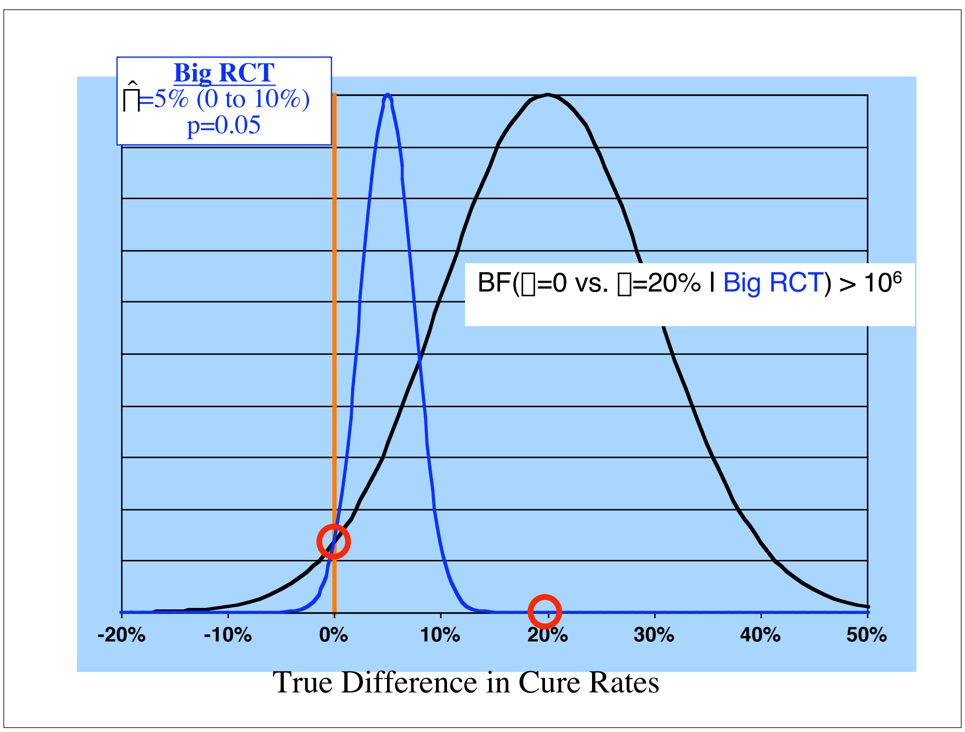
$$\frac{L(\Delta = 0 \mid \bar{x})}{Max_{\Delta} L(\Delta \mid \bar{x})} = e^{-Z^2/2}$$

- The ratio of the data's probability under the null hypothesis versus the hypothesis that the observed effect is the true one.
- The smallest possible likelihood ratio (or Bayes Factor) for the null hypothesis versus any other hypothesis.

S. Goodman, FDA Bayes Workshop




Small RCT
$\hat{□}$=20% (0 to 40%)
p=0.05
BF(□=0 vs. □=5% | Small RCT) = 0.4
-20%
-10%
0%
10%
20%
30%
40%
50%
True Difference in Cure Rates

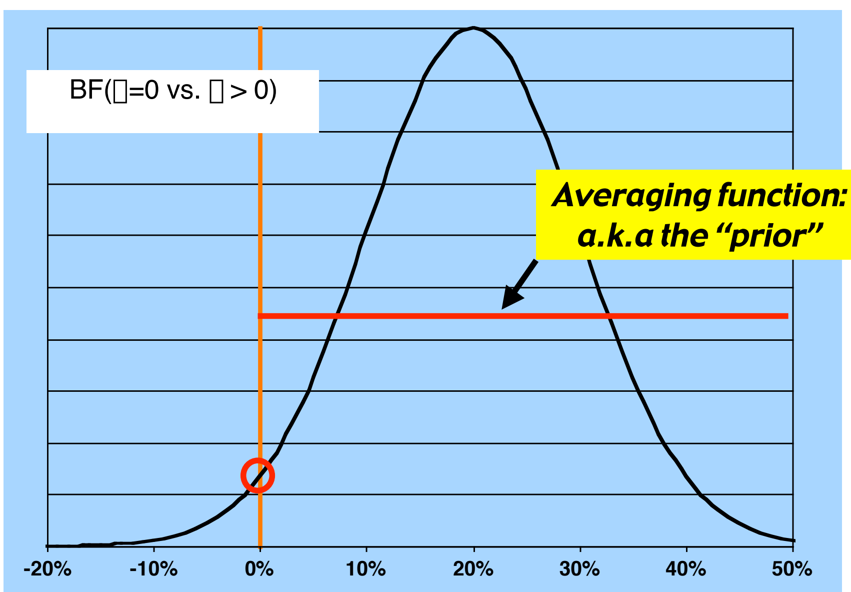

Big RCT
$\hat{□}$=5% (0 to 10%)
p=0.05
BF(□=0 vs. □=20% | Big RCT) > 10^6
-20%
-10%
0%
10%
20%
30%
40%
50%
True Difference in Cure Rates

Dependence of Evidence on Alternative Hypothesis

Alternative Hypothesis	Data (P=0.05)	BF (H_0 vs. H_1)
□ = 5%	Big Trial (5%)	0.14
□ = 20%	Small Trial (20%)	0.14
□ = 5%	Small Trial (20%)	0.4
□ = 20%	Big Trial (5%)	> 10^6

S. Goodman, FDA Bayes Workshop

Averaging the likelihood: The Bayes Factor



True Difference in Cure Rates

P-values: Bayesian Translations

P-value *(Z-score)*	Minimum Bayes factor	-e p ln(p)	Strength of evidence	Decrease in probability of the null hypothesis, % From	To no less than
0.10 *(1.64)*	.26	.6	Weak	75 50 17	44 21 5
0.05 *(1.96)*	.15	.4	Moderate	75 50 26	31 13 5
0.03 *(2.17)*	.1	.3	Moderate	75 50 33	22 9 5
0.01 *(2.58)*	.04	.13	Moderate to strong	75 50 60	10 3.5 5
0.001 *(3.28)*	.005	.02	Strong to very strong	75 50 92	1 0.5 5

Stopping Rules

Stopping Rule "Paradox"

The probability of misleading evidence (small p-value) approaches 100% as # of looks -> .

$\Pr(p < \square)\ \mathrm{l}\ H_0)\ \square\square\square > 1$

But!!

$\Pr(BF < \square \mid H_0) \leq \square$

The Type I error rate has a relationship to evidential strength, *but only when the evidence is measured properly.*

P-values = Data dredging

- The high Type I error rate with multiple looks is created by is produced when we summarize the likelihood curve at a point *determined by the data.*
- Using p-values is like data dredging, in that we measure the evidence for a data-suggested hypothesis instead of averaging the evidence over pre-specified simple hypotheses.
- The optimal averaging is done with a Bayesian prior. ***This is why Bayesian methods can be viewed as a "calculus of evidence" as well as a "calculus of belief."***

What FDA Needs to Know About Bayesian Statistics

- That Bayes theorem has separable data and belief components, and can be viewed as a calculus of evidence, not just belief.
- That likelihood-based evidential measures have very attractive frequentist properties, as well as a sound theoretical foundation and intuitive interpretations.
- That standard inferential methods represent evidence inappropriately, and produce unnecessary rigidity in design and interpretation.

And...

- That the use of Bayesian evidential measures can have an impact far beyond the (sometimes) different numbers they produce; they can affect how we talk about evidence, and who participates in that dialogue.

S. Goodman, FDA Bayes Workshop

Final thoughts

"What used to be called judgment is now called prejudice, and what used to be called prejudice is now called the null hypothesis....it is dangerous nonsense (dressed up as 'the scientific method') and will cause much trouble before it is widely appreciated as such."

A.W.F. Edwards (1972)

S. Goodman, FDA Bayes Workshop