Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

Likelihood Ratios and the Strength of Statistical Evidence

Jeffrey D. Blume
Center for Statistical Sciences
Brown University, Providence, RI

Three questions

These generic questions define three distinct problem-areas of statistics:
1. What should I believe? (Bayesian Inference)
2. What should I do? (Decision Theory)
3. What do these data say? (Evidential Analysis)

Hypothesis testing

- Neyman and Pearson (1933)
- "The problem of testing statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or take step B..."
- "Thus to accept a hypothesis means only to decide to take action A rather than action B; this does not mean that we necessarily believe that the hypothesis H is true;" (Neyman 1950)
Hypothesis testing

- Concept of controlling type I and II errors is attractive; Leads to 'inductive behavior'

- No concept of statistical evidence

Significance testing

- ‘P-values measure the strength of evidence against a hypothesis’ (Fisher 1958)

- Interpretation of p-value depends on sample size
  1. Berkson (1942), Cornfield (1966)
  2. Lindley & Scott (1984), Peto et al. (1976)

Significance testing

- Over time p-values have been given a post-hoc type I error interpretation, leading to irresolvable controversies over multiple comparisons and multiple looks

- Why?
  Because p-values represent two distinct concepts:
  1. The measure of the strength of evidence
  2. The measure of the potential for that evidence to be misleading
Bayesian inference

- Rev. T. Bayes (1763), Savage (1954), Jeffreys (61), Lindley (65), De Finetti (74)
- "The key ideas of Bayesian statistics [are] that probability is orderly opinion, and that inference from data is nothing other than revision of such opinion in the light of relevant new information;"

Bayesian inference

- "[The Bayesian approach] is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data;"
  (Edwards, Lindman & Savage, 1963)
- Priors, Posteriors, & Bayes factors
- Do uninformative priors exist?
  (definition, reparameterization)

A diagnostic test

<table>
<thead>
<tr>
<th></th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive</td>
</tr>
<tr>
<td>Disease yes</td>
<td>0.94</td>
</tr>
<tr>
<td>Status no</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Suppose we observe a positive test result.
Diagnostic test

A physician might ask:
1. Should this observation lead me to believe that this person has the disease?
2. Does this observation justify my acting as if he has the disease?
3. Is this test result evidence that he has the disease?

The law of likelihood

If hypothesis A implies that the probability of observing some data $X$ is $P_A(X)$, while hypothesis $B$ implies that the probability is $P_B(X)$, then the observation $X = x$ is evidence supporting $A$ over $B$ if $P_A(x) > P_B(x)$, and the likelihood ratio, $P_A(x)/P_B(x)$, measures the strength of that evidence;
(Hacking 1965, Royall 1997)

The law says

- The strength of evidence is measured by the likelihood ratio: $LR = P_A(x)/P_B(x)$
- "$H_A$ is supported over $H_B$ by a factor of $LR$;"
  - If $LR = 1$, the evidence is neutral
  - If $LR > 1$, the evidence supports $H_A$ over $H_B$
  - If $LR < 1$, the evidence supports $H_B$ over $H_A$
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

Degrees of strength

- For interpreting and communicating the strength of evidence, it is useful to divide the scale into descriptive categories
- **Benchmarks** of LR = 8 and 32 are used to guide that description

• **Weak evidence**
  - For $H_A$ over $H_B$: $1 < LR < 8$
  - For $H_B$ over $H_A$: $1/8 < LR < 1$

• **Moderate evidence**
  - For $H_A$ over $H_B$: $8 < LR < 32$
  - For $H_B$ over $H_A$: $1/32 < LR < 1/8$

• **Strong evidence**
  - For $H_A$ over $H_B$: $32 < LR$
  - For $H_B$ over $H_A$: $LR < 1/32$

Diagnostic test, revisited

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>yes</td>
<td>0.94</td>
</tr>
<tr>
<td>negative</td>
<td>no</td>
<td>0.02</td>
</tr>
</tbody>
</table>

A positive test result is statistical evidence supporting $H_B$, over $H_A$. (b/c 0.94 > 0.02)
\[ LR = \frac{P(T^+|H_D)}{P(T^+|H_{\sim D})} = \frac{0.94}{0.02} = 47 \]

- Because

\[ P(T^+|H_D) = 0.94 \]

- The evidence itself can be misleading, but here misleading positive results are rare;

\[ P(T^+|H_{\sim D}) = 0.02 \]

Irrelevant for data interpretation

- Once the observations have been collected, the design probabilities are irrelevant; they play no role in the analysis stage;

- They should only be considered at the planning stage;

The diagnostic test

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.94</td>
</tr>
<tr>
<td>no</td>
<td>0.02</td>
</tr>
</tbody>
</table>

\[ LR = \frac{P(T^+|H_D)}{P(T^+|H_{\sim D})} = \frac{0.94}{0.02} = 47 \]

\[ P(T^+|H_D) = 0.02 \]
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

Another diagnostic test

<table>
<thead>
<tr>
<th>Test #2</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive</td>
</tr>
<tr>
<td>Disease</td>
<td>0.47</td>
</tr>
<tr>
<td>Status</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[ LR = \frac{P(T^+|H_1)}{P(T^+|H_0)} = \frac{0.47}{0.01} = 47 \]

\[ P(T^+|H_0) = 0.01 \]

Question

- Is a positive result on the second test stronger evidence in favor of disease than a positive result on the first one?
- Is the positive result on the second test "less likely to be misleading", "more reliable" in some sense, or does it warrant more "confidence"?

Answer

- **No**: A positive result on the second test is equivalent, as evidence about the presence or absence of disease, to a positive result on the first.
Why?

- An observed positive result is misleading if and only if the subject does not have the disease and \( P(\text{no disease}|T^+ \) is the same for both tests!

\[
P(D^-|T^+) = \frac{P(T^+|D^-)P(D^-)}{P(T^+|D^-)P(D^-) + P(T^+|D^+)P(D^+)}
\]

\[
= \frac{P(D^-)}{P(D^-)+\frac{P(T^+|D^+)P(D^+)}{P(T^+|D^-)P(D^-)}} = \frac{P(D^-)}{P(D^-)+4P(D^+)}
\]

- An observed positive result is no more likely to be misleading if it comes from one test than if it comes from the other.

Three key concepts

1. A measure of the strength of the evidence
2. The probability that a particular design will generate misleading evidence
3. The probability that observed evidence is misleading
4. Current frequentist and Bayesian methods fail to distinguish between these concepts
Three key concepts

- While the probability in (2) depends on the number of looks at the data, the probability in (3) does not;
- Failure to make this distinction has led to the multiple looks controversy over adjustments of tail area probabilities;

Old news

- “In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence;” [Fisher, 1959]

The University Group Diabetes Program

- Multi-centered, randomized clinical trial, to evaluate the effect of Tolbutamide on vascular complications of adult-onset diabetes (1961-1975);
- Is there evidence that Tolbutamide increases the probability of cardiovascular death?
- Likelihood analysis by Blume (Stat. Med., 2002);
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

UGDP data

<table>
<thead>
<tr>
<th>Center</th>
<th>Tolbutamide Deaths</th>
<th>Placebo Deaths</th>
<th>Total Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baltimore</td>
<td>1</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>7</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Cleveland</td>
<td>1</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>6</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>New York</td>
<td>2</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>Williamson</td>
<td>3</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Birmingham</td>
<td>2</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Boston</td>
<td>4</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Chicago</td>
<td>6</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>St. Louis</td>
<td>8</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>San Juan</td>
<td>6</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Seattle</td>
<td>0</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>All Centers</td>
<td>26</td>
<td>204</td>
<td>205</td>
</tr>
</tbody>
</table>

Binomial likelihood

- Θ is the probability of placebo CV death

\[ P(s = 10 \mid \theta, n = 205) \propto \theta^{10}(1 - \theta)^{205 - 10} = L(\theta) \]

- The evidence for Θ = 0.1 versus Θ = 0.05 is measured by

\[ L(0.1)/L(0.05) = 1/37 \]
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

Probability of cardiovascular death
Tolbutamide group

Relative risk of cardiovascular death
Tolbutamide versus placebo

Relative risk of cardiovascular death

The screen versions of these slides have full details of copyright and acknowledgements
‘Undesirable’ evidence

1. Weak evidence (1/8 < LR < 8)
2. Misleading evidence: strong evidence in favor of an incorrect hypothesis over the correct hypothesis

Example: $H_0: X \sim g$  $H_0: X \sim \hat{f}$

If $X_1,\ldots,X_n$ are i.i.d. $g$ and $LR = \prod_{i=1}^{n} \frac{f(x_i)}{g(x_i)} = 40$

Design considerations

- The probabilities of observing misleading and weak evidence provide quantities analogous to the type I and II error rates

Example: Collect $X_1,\ldots,X_n \sim$ i.i.d. $\text{N}(\mu, \sigma^2)$

$H_0: \mu = \mu_0$, $H_1: \mu = \mu_1$, $LR = \frac{L(\mu_1)}{L(\mu_0)}$,

$\Delta = (\mu_1 - \mu_0)/\sigma$

• Then the probabilities of observing misleading and weak evidence are

$$M(n, k) = P \left( \frac{L(\mu_1)}{L(\mu_0)} \geq k \right) = \Phi \left( -\frac{\sqrt{n} \Delta \mu_1}{\Delta n} - \frac{\ln k}{\Delta n} \right)$$

$$W(n, k) = P \left( \frac{1}{k} \frac{L(\mu_1)}{L(\mu_0)} < k \right) = \Phi \left( \frac{\sqrt{n} \Delta \mu_1}{\Delta n} \right) \Phi \left( \frac{\ln k}{\Delta n} \right)$$
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

Efficiency of sequential designs
• The likelihood function is unaffected by the number of examinations; hence re-examination of accumulating evidence does not diminish its strength
• Maximizes flexibility and participant safety
• Sequential designs require substantially fewer participants than their fixed sample size counterparts

Simulation
1. Generate $X_1,X_2,X_3,\ldots \sim \text{i.i.d. } N(\mu_1,\sigma^2), \Delta = 1$
2. Examine the data after each observation; stopping only if $L(\mu_1)/L(\mu_0) > k$ or $L(\mu_0)/L(\mu_1) > k$
3. Repeat 100,000 times

The screen versions of these slides have full details of copyright and acknowledgements
• Average sample sizes for sequential designs with these operating characteristics are 50% smaller; Note that these savings are invariant to \( \theta \) and are well known (Wald, 1947)

### Results

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misleading</td>
<td>Avg. 25% 50% 75% Max.</td>
</tr>
<tr>
<td>K = 8</td>
<td>0.000 0.054 4.39 2 4 6 48</td>
</tr>
<tr>
<td>K = 10</td>
<td>0.000 0.003 4.47 2 4 6 48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type I Power Fixed sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>N, P</td>
</tr>
<tr>
<td>0.05</td>
</tr>
<tr>
<td>0.50</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

The universal bound

• For any pair of probability distributions, the probability of observing misleading evidence of at least \( k \)-strength is bounded by \( 1/k \)

\[
\text{for all } n, \quad P \left( \prod_{i=1}^{n} \frac{f(X_i)}{g(X_i)} \geq k \right) \leq \frac{1}{k}
\]

• The universal bound holds for sequential designs as well

\[
P \left( \prod_{i=1}^{n} \frac{f(X_i)}{g(X_i)} \geq k \right) \leq \frac{1}{k} \quad \text{for any } n = 1, 2, \ldots
\]

• This is a very strong property: it is not possible to frequently observe misleading evidence even if you are doing your best to generate it
The bump function

- Represents the probability of observing misleading evidence in fixed sample size designs, as a function of the alternative
- \(X_1, \ldots, X_n \sim \text{i.i.d. } N(\mu_0, \sigma^2); \; \mu \text{ fixed}; \; S_n = X_1 + \ldots + X_n\)
- Measure the evidence for
  - \(H_0: \mu = \mu_0 \text{ vs. } H_0: \mu = \mu_0; \; \Delta = |\mu - \mu_0|/\sigma\)

Bump equations

\[
M(n, k) = P \left( \frac{L(\mu_1)}{L(\mu_0)} \geq k \right) = P \left( S_n = \frac{\Delta}{\sigma} n + \frac{\ln k}{\Delta} \right) = \Phi \left( \frac{-\Delta n}{2} \frac{\ln k}{\Delta n} \right) = \Phi \left( -\frac{c}{2} \frac{\ln k}{c} \right)
\]

- Maximum is \(\Phi \left( -\sqrt{2 \ln k} \right) = 0.021 (k = 8)\)

The screen versions of these slides have full details of copyright and acknowledgements
The tepee function

- Represents the probability of generating misleading evidence for $H_a$ over $H_0$
  in an open sequential design, under a normal model
- Examine the data after each observation;
  stopping only when the likelihood ratio for $H_a$
  versus $H_0$ is greater than or equal to $k$

Tepee equations

$$P_0\left(\frac{L(\mu_1)}{L(\mu_0)} \geq k \right) ; \text{for any } n = 1, 2, ...$$

$$= P_0\left(S_n \geq \frac{\Delta}{\sigma} n + \frac{\ln(k)}{\Delta} ; \text{for any } n = 1, 2, ... \right) \geq \frac{1}{k} \exp\left(-\rho \Delta\right)$$

- $\rho$ is the expected overshoot

The probability of generating misleading evidence

The tepee function
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

The probability of generating misleading evidence

Composite hypotheses

- The law of likelihood explains that statistical evidence depends on two simple hypotheses
- For this reason the law implies that there can be no unique measure of evidence when a composite hypothesis is involved (Analogy: statistical power)

Composite hypotheses

- To convey the strength of evidence over the parameter space simply draw the likelihood function!
- That is, show all the likelihood ratios
Relative risk of cardiovascular death

So what?

- Provides a practical frequentist justification of the law of likelihood
- Demonstrates that the act of continuously monitoring a trial only minimally increases the chances of observing misleading evidence

Extensions

The bump and tepee functions represent the limiting probability when
1. The underlying distribution is non-normal with a single parameter of interest
2. Fixed dimensional nuisance parameters are present and a profile likelihood is used
3. When robust likelihood functions are used (Royall and Tsou 2003; Blume et al., 2006)
Likelihood Ratios and the Strength of Statistical Evidence
Dr. Jeffrey D. Blume

References
- Royall, 1997; Statistical Evidence; Chapman & Hall, London
- Blume, Su, Olveda, McGarvey, 2006; Statistical Evidence for GLM Regression Parameters: A Robust Likelihood Approach; To appear in Statistics in Medicine