Acknowledgements

• Scott Zeger

• Marie Diener-West

• ICTR Leadership / Team

Introduction to Statistical Inference

1. The game of statistical inference
2. Three key questions – three approaches to inference: likelihood; Bayesian; frequentist
3. Frequentist statistics
   • An overview: Sampling variability
   • Tests of hypotheses – p-values
   • Uncertainty in estimation – confidence intervals
4. Brief introduction to data scales
What Is Biostatistics?

The science of learning from biomedical data involving appreciable variability or uncertainty

Game of Statistical Inference: “Search for Truth”

Example: Risperidone versus Haloperidol for Prevention of Schizophrenic Symptoms

• Randomized trial comparing risperidone to haloperidol for time-to-relapse in schizophrenic patients

• Reference: Csernansky, et al. 2002. NEJM

• 367 patients followed for up to 800 days

• Hypothesis: relative to haloperidol, risperidone will prolong the time to relapse of schizophrenic symptoms

• Focus on binary indicator of relapse or not at 400 days for now
Statistical Inference

An approach for using information about the sample to make statements about the population

Which population?

• The one from which your sample was randomly drawn
• The one for which your sample is reasonably thought to be representative.

Three Questions in Clinical Research

• What is the evidence in the data?
  – Likelihood function measures evidence

• What do we believe in light of the evidence?
  – Combine prior with likelihood to obtain posterior distribution

• What should we decide to do in light of the evidence?
  – Frequentist statistics for decision rules
  – Controlling types of errors (rate of wrong decisions)

Three Questions: The Example

• What is the evidence from the Czernansky data that risperidone prolongs time to relapse relative to haloperidol?

• What do we believe about this difference in light of this evidence?

• Should the FDA give J&J an “indication” that risperidone is better than haloperidol at delaying time to relapse?
Evidence
• Measured by the “likelihood function”

• Is always relative: supporting one hypothesis relative to another

• Is what science generates

• Is used to update prior beliefs to address question 2

Data = Evidence

2x2 Table – Relapse within 400 days; from Figure 1 in Csernansky, et al. 2002

<table>
<thead>
<tr>
<th></th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relapse</td>
<td>41</td>
<td>67</td>
<td>108</td>
</tr>
<tr>
<td>No Relapse</td>
<td>136</td>
<td>121</td>
<td>257</td>
</tr>
<tr>
<td>Total</td>
<td>177</td>
<td>188</td>
<td>365</td>
</tr>
</tbody>
</table>

Observed Relative Rate of Relapse

• Risperidone: Relapse rate = 41/177 = 0.23

• Haloperidol: Relapse rate = 67/188 = 0.35

• Relative rate = Risp rate/Hal rate = 41/177 / 67/188 = 0.23/0.35 = 0.65
Likelihood Function for Relative Rate of Relapse

Truth: rate of relapse for haloperidol; relative rate for risperidone versus haloperidol

Probability model: an equation describing the probability of observing particular rates in one’s sample if their population has certain characteristics

Statistical inference: use data above to make statements about true relative rate

Likelihood Function is the Probability of the Observed Data as a Function of True Parameters = Rate for halo; relative rate

\[
p = \text{rate for halo} \\
rr = \text{relative rate}
\]

The likelihood function of the data is:

\[
(rr \times p)^{41}(1 - rr \times p)^{136} p^{67} (1 - p)^{121}
\]

Likelihood as a function of RR

Maximum Likelihood Estimate
Likelihood as a function of RR

Relapse Data – Half and Twice the Sample Size

<table>
<thead>
<tr>
<th></th>
<th>Half</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risp</td>
<td>Hal</td>
<td>Total</td>
<td>Relapse</td>
<td>Risp</td>
<td>Hal</td>
<td>Total</td>
<td>Relapse</td>
<td>Risp</td>
<td>Hal</td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relapse</td>
<td>21</td>
<td>34</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Relapse</td>
<td>68</td>
<td>60</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>94</td>
<td>182</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
<th>Relapse</th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
<th>Relapse</th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Relapse</td>
<td>82</td>
<td>154</td>
<td>216</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Relapse</td>
<td>272</td>
<td>242</td>
<td>514</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td>376</td>
<td>730</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Likelihood Functions and Sample Size
Notes on Likelihood Function

"Best estimate" of RR is the one that maximizes the (profile) likelihood function – “maximum likelihood estimate” or mle

• The “spread” of the likelihood function reflects the strength of evidence – more concentrated means stronger evidence

• The more data, the more concentrated the likelihood function becomes around the true value

Summary of Evidence

• Value for the true, but unknown relative risk that makes the observed data most likely is 0.65, that is 35% fewer relapses for patients on risperidone

• A range of values for the true relative risk that make the data more than 1/32 times as likely as the best value (0.65) is (0.41 to 1.00).

• The data are 50 times more likely if the relative risk is 0.65 than if it is 1.0.

Bayesian Methods

• On the ascendance – beginning to be seen at FDA

• Treat unknowns (true relative rate) as random and give probability distributions as measures of belief

• More in future courses
Frequentist Inference

- **Big idea:** imagine what data might have occurred if a particular hypothesis, H, were true ("under H")

- Example: assume risperidone is identical to haloperidol in rate of relapse
  - Relative rate = 1.0

- Calculate the distribution of sample relative rates under this assumption

- Compare what actually occurred with what tends to occur when relative rate is 1.0

- If what occurred is rare under H, "reject" H

---

Procedure

Play Supreme Ruler: create a world in which relative rate is really 1.0

Use computer to generate lots of data sets with same numbers of persons per group and same average rate of relapse as observed in our data (108/365); calculate sample relative rate for each

Make a distribution of relative rate estimates over all data sets: "null distribution" for relative rate estimate

---

First Simulated Data Set

<table>
<thead>
<tr>
<th></th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relapse</td>
<td>56</td>
<td>52</td>
<td>108</td>
</tr>
<tr>
<td>No Relapse</td>
<td>121</td>
<td>136</td>
<td>257</td>
</tr>
<tr>
<td>Total</td>
<td>177</td>
<td>188</td>
<td>365</td>
</tr>
</tbody>
</table>

Relative risk for this data: 56/177 / 52/188 = 1.14
Repeat 1000 Times

Relative Rates:

- 1.44
- 1.02
- 0.76
- 1.19
- 0.98
- 0.88

What Have We Learned?

- We rarely get observed relative rates as small as 0.65 in a study the size of this one when the truth is that risperidone is not different than haloperidol (true relative rate = 1.0)
  - Here, as small or smaller than 0.65 ~ 1.8% of samples

- We therefore “reject” the null hypothesis that the true relative rate is 1.0 in favor of the alternative hypothesis that it is less than 1.0

What Have We Learned?

- When the relative risk=1.0 (null hypothesis is true), the probability of observing a sample relative rate of 0.65 or smaller is 0.018

- When the relative risk=1.0 (null hypothesis is true), the probability of observing a sample relative rate of 0.65 or smaller, or 1.0/0.65=1.54 or bigger, is 0.035

- Two-sided p-value = 0.035
**p-value**

- The probability of observing a test statistic as or more extreme than occurred in sample under the null hypothesis

- **Probability:** How often as we imagine what other data might have occurred in our sample under particular hypotheses

- **As or more extreme:** as different or more different from the hypothesized value

  - This is "two sided"; if “one sided”, “probability of ... a ... statistic larger than..."

**under the null hypothesis:** If the sample truly comes from the population we hypothesize
p-value: “p” is for:

Publish

Perish if you don’t publish

aPProval of your drug by FDA

Promotion to Vice-President for Research

Point-05 = 0.05 (or 0.01)… common “cutoffs” to reject

Hypothesis Testing

• Assume a null hypothesis – straw man to knock down (no difference between risp and hal or \( rr=1 \))

• Choose a test statistic (relative risk)

• Compute the null distribution (here, by simulation; often, by probability model)

• Calculate p-value

• Reject the null if p-value less than a selected (“alpha”) level – usually 0.05

Repeat 1000 Times

<table>
<thead>
<tr>
<th>Relative Rates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.14</td>
</tr>
<tr>
<td>1.02</td>
</tr>
<tr>
<td>0.76</td>
</tr>
<tr>
<td>1.19</td>
</tr>
<tr>
<td>0.98</td>
</tr>
<tr>
<td>0.88</td>
</tr>
</tbody>
</table>

Note: The histogram shows the distribution of relative rates from 1000 simulations.
Two-sided p-value = 0.035 < 0.05; conclude that the relapse rates are not equivalent for the two drugs (reject the "null" hypothesis that they are).

A name for the "distribution" of values that might have occurred:

- **SAMPLING DISTRIBUTION**
  - These can be computed "under the null hypothesis" or under any other hypothesis
  - There is a way to approximate it for one's data— that is, for the population truly underlying one's sample: "Bootstrapping"

Another Frequentist Method: "Bootstrapping"

- Imagine that the sample is the whole population
- Repeat the experiment a large number of times to obtain a distribution for the statistic of interest
  - Draw a bootstrap sample of 177 risp and 188 hal patients from the original sample, now the population
  - With replacement
  - Calculate the relative risk
  - Repeat 1000s of times
- Find a central interval that includes 95% of relative risks: "confidence interval"
First Simulated Data Set

<table>
<thead>
<tr>
<th></th>
<th>Risp</th>
<th>Hal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relapse</td>
<td>44</td>
<td>54</td>
<td>98</td>
</tr>
<tr>
<td>No Relapse</td>
<td>133</td>
<td>134</td>
<td>267</td>
</tr>
<tr>
<td>Total</td>
<td>177</td>
<td>188</td>
<td>365</td>
</tr>
</tbody>
</table>

Relative risk for this data: \( \frac{44}{177} / \frac{54}{188} = 0.87 \)

Repeat 1000 Times

Relative Rate:
- 0.87
- 0.73
- 0.53
- 0.53
- 0.66
- ...
- 0.80

Confidence Interval for True Relative Risk

We are 95% confident that the interval: 0.45, 0.89 includes the true, unknown value
(95% of 95% confidence intervals contain the truth)
Confidence interval – Main Idea

Confidence Interval for True Relative Risk

We are 95% confident that the interval: 0.45, 0.89 includes the true, unknown value
(95% of 95% confidence intervals contain the truth)

In frequentist statistics, the sample and hence particular interval is random. The truth is fixed

We don’t say: “the probability that the true value is between 0.45 to 0.89 is 0.95.”

Scales and Types of Data

• Scales
  – Numeric
    • Continuous – no gaps in the possible measurements
    • Otherwise: discrete (e.g. counts)
  – Categorical
    • Binary, dichotomous – two categories
    • Polytomous – more than two
    • Ordinal – ordered categories (e.g. satisfaction ratings)
Scales and Types of Data

- A few important categories of data
  - “Measures” – determined relatively precisely
  - Constructs – measurable only up to considerable error
  - Times-to-events – may be “censored”—only known to be later (or earlier) than a certain point: later in course
  - Variable – label for one aspect of data collected

- Why this all matters
  
  Data scale and type partially determine the correct analysis to use!

Time to Event Outcomes

Main Points Once Again - Inference

- The game of statistical inference is to use data to make statements about how the world works

- Three key questions:
  1. What is the evidence – likelihood function;
  2. What do we believe? – Bayesian methods
  3. What do we decide - frequentist methods

- Frequentist statistics
  - Tests of hypotheses – p-values
  - Confidence intervals (via bootstrapping or model)