Topics in Dichotomous Test Evaluation

Henry Glick

Epi 550

January 21, 2015

WHY DO WE TEST?

Why Do We Test?

1) To increase or decrease likelihood of disease so we are sufficiently confident in treating or withholding treatment
2) To understand disease process simply for sake of knowing
3) To avoid malpractice
4) To generate revenue
WHAT DO WE MEAN WHEN WE SAY THAT A TEST RESULT IS “POSITIVE”?

What Do We Mean When We Say That a Test Result is “Positive”?

1) Person has disease?
2) Test result more than 2 sd above (or below) mean?
3) Test result leads to a post-test (or posterior) probability that is greater than pre-test (or prior) probability?
4) If we don’t plan to obtain additional information, test result is indicative of treatment?
   ?? “Test result was positive, but we are going to send you home anyway” ??

DIAGNOSTIC TESTING OUTLINE
(the next 4 weeks)

- Interpreting dichotomous tests (FINISHING TODAY)
  - 2x2 tables
  - Likelihood ratios positive and negative
- Interpreting continuously scaled tests
  - Selection of optimal 2x2 table (with and without use of receiver operating characteristic (ROC) curves)
  - Stratum-specific likelihood ratios (SSLR)
- (Break) Development of prediction rules
- Verification bias
- Comparison of 2x2 and SSLR approaches
- Choice among tests
- Stability of test results
Outline for Today

- Treatment threshold and costs of mistakes
- LR+/- and sensitivity and specificity
- Relationship between likelihood ratios, odds ratios, and relative risks
- Confidence intervals for test characteristics
- Sample size for determination of test characteristics
- Probabilists vs decision makers

1. Treatment Threshold and Cost of Mistakes

- Last class, Sankey identified two common treatment thresholds that exist when a diagnostic test is available: do nothing / test threshold and test / treat threshold

<table>
<thead>
<tr>
<th>No test- No treat</th>
<th>Test and Treat if Test result is positive</th>
<th>Treat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Probability of disease ^TT</td>
<td>1</td>
</tr>
<tr>
<td>^TT</td>
<td>^TT TTT</td>
<td></td>
</tr>
</tbody>
</table>

Other Thresholds

- Other thresholds exist
  - e.g., test and treat; withdraw treatment if test result is negative
  - Used when expected cost of testing, treating, and withdrawing treatment after a negative test is less than expected cost of testing and initiating treatment if test is positive
What If No Diagnostic Test Is Available?

- Suppose you are in a remote health center and a child aged between 3 and 36 months presents with a rectal temperature >39°C
- Child has no obvious focal infection for which timely antibiotic therapy is indicated (e.g., otitis media), nor has she received antibiotics during preceding 48 hours
- She has no “toxic” clinical appearance necessitating immediate hospitalization, nor a specific viral infection (e.g., varicella), a known immune-deficiency condition, or chronic illness that would alter standard approaches to febrile illness (e.g., hemoglobinopathy)
- What information goes into decision to either treat empirically or watchfully wait?

Likelihood (Probability) of Disease

- All else equal, higher the likelihood (probability) of disease, more likely empirical treatment indicated
- Lower the likelihood, less likely empirical treatment indicated

Cost of Mistakes

- All else equal, higher the cost of mistakenly withholding treatment compared with mistakenly treating, more likely empirical treat indicated
- Lower the cost of mistakenly withholding treatment compared with mistakenly treating less likely empirical treatment indicated
Treatment Threshold

- Information about probability of disease and cost of mistakes can be combined to identify a probability of disease \((p^*)\) where expected costs of mistakenly treating and mistakenly withholding treatment are equal
  - If probability of disease is above \(p^*\), cost of mistakes from withholding treatment exceed those of mistakenly treating, and treat indicated
  - If it is below \(p^*\), reverse is true and withholding treatment indicated

Definitions

\(OTP\) = Value of outcome given true positive
\(OFN\) = Value of outcome given false negative
\(OTN\) = Value of outcome given true negative
\(OFP\) = Value of outcome given false positive
\(OTP - OFN = CFN\) = Cost of false negative
\(OTN - OFP = CFP\) = Cost of false positive

Expected outcome of treatment (\(EOTreat\))
\(EOTreat = pOTP + (1-p)O FP\)

Expected outcome of no treatment (\(EONoTreat\))
\(EONoTreat = pOFN + (1-p)OTN\)

Treatment threshold = \(p^* = (EOTreat = EONoTreat)\)

Deriving Treatment Threshold

Solve for \(p^*\) such that \((EOTreat = EONoTreat)\)

\[1\]
\[pOTP + (1-p)O FP = pOFN + (1-p)OTN\]

\[2\]
\[pOTP + O FP - pO FP = pOFN + OTN - pOTN\]

\[3\]
\[(pOTP - pOFN) + (pOFN - pO FP) = (OTN - O FP)\]

\[4\]
\[p[OtP - O FN] + (O FN - O FP)] = (OTN - O FP)\]

\[5\]
\[p(OFN + O FP) = CFP\]

\[6\]
\[p^* = CFP / (CFN + CFP)\]
Treatment Decisions and Threshold

\[ \frac{C_{FP}}{C_{FN} + C_{FP}} = \text{Treatment threshold} \]

- At end of any testing sequence (e.g., no tests or 1, 2, 3+ tests):
  - If (posttest) probability is less than threshold, watchfully wait because expected costs of mistakenly treating exceed those of mistakenly withholding treatment
  - If probability is greater than threshold, treat empirically because expected costs of mistaken withholding of treatment exceed those of mistaken treatment

Treatment Threshold and Definition of Positive and Negative Tests

- Treatment threshold is (meant to be) linked to definition of positive and negative tests
  - Independent of whether test result leads to a post-test probability that exceeds pretest probability, negative tests should yield post test probabilities that are below treatment threshold
  - Independent of whether test result leads to a post-test probability that is less than pretest probability, positive tests should yield post test probabilities that exceed treatment threshold

Definition of Costs of Mistakes

- Difference in net value of treating someone correctly and net value of treating them incorrectly
- Can be estimated by use of a cost-benefit framework (monetizing both costs and outcomes)
- Also can be estimated by use of a cost-effectiveness (NMB) framework (separate calculation of costs (c) and outcomes (e)):
  - \[ C_{FN} = (W \cdot \theta_{FN}) - c_{FN} \]
  - \[ C_{FP} = (W \cdot \theta_{FP}) - c_{FP} \]
  - Where \( W \) = willingness to pay
Cost-Effectiveness Equivalent

- Under a cost-effectiveness framework (in which $C_{FN}$ equals a combination of $c_{FN}$ and $e_{FN}$ and $C_{FP}$ equals a combination of $c_{FP}$ and $e_{FP}$):

$$p^* = \frac{W(e_{FP} - c_{FP})}{W(e_{FP} + e_{FN}) - (c_{FP} + c_{FN})}$$

Ratio of Costs Sufficient

- Many people uncomfortable with identifying costs of false positive and false negative decisions, particularly their absolute magnitudes
- Good news: Ratio of costs more important than absolute magnitudes
  - e.g., when defining treatment threshold, if cost of false positives thought to be 1/3 the cost of false negatives, know that:

  $$C_{FP} = \frac{1}{3} C_{FN}$$

  $$\Rightarrow p^* = \frac{1}{3} C_{FN} / (\frac{1}{3} C_{FN} + C_{FN}) = 0.25$$

Treatment Threshold and Costs

- If instead have an idea of treatment threshold, can infer relative valuation of $C_{FP}$ and $C_{FN}$
  - If our $p^* = 0.25$, then

  $$0.25 = \frac{C_{FP}}{(C_{FN} + C_{FP})}$$

  $$0.25C_{FN} + 0.25C_{FP} = C_{FP}$$

  $$0.25C_{FN} = 0.75 C_{FP}$$

  $$\frac{1}{3} C_{FN} = C_{FP}$$

- Implication: Implicitly make assumptions about $C_{FP}$ and $C_{FN}$ whenever a treatment decision is made
2. LR+/- and Sensitivity and Specificity

- Like sensitivity and specificity, likelihood ratios are characteristics of test itself
- Combined with data on pretest probability of disease [or a transformation of this probability such as prior odds] to obtain a post test probability of disease
- If sensitivity and specificity are independent of prevalence, likelihood ratios also independent of prevalence
- If test result is truly dichotomous, no advantages to use of sensitivity and specificity vs likelihood ratios for positive and negative tests
  - Corollary: when test result is continuous, information from likelihood ratios can differ from information from sensitivity and specificity

3. What Is Relationship Between Likelihood Ratios, Odds Ratios, and Relative Risks?

- Likelihood ratios, odds ratios, and relative risks share certain features, including that all 3 are commonly calculated by use of 2x2 tables, but they also differ

<table>
<thead>
<tr>
<th>Disease</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result positive</td>
<td>(a) True positive (TP)</td>
<td>(b) False positive (FP)</td>
<td>a+b = g</td>
</tr>
<tr>
<td>Test result Negative</td>
<td>(c) False negative (FN)</td>
<td>(d) True negative (TN)</td>
<td>c+d = h</td>
</tr>
<tr>
<td>Total</td>
<td>All with disease</td>
<td>All normals</td>
<td>(a+c) = e</td>
</tr>
</tbody>
</table>
Likelihood Ratio and Odds Ratio Similarities

• Both refer to relative frequency of an outcome
  – Likelihood ratios: relative frequency of test result among those with and without disease
  – Odds ratios: relative frequency of disease among those who are and who are not exposed to a risk factor for disease
• Both can be used to calculate posterior probabilities of disease

Likelihood Ratio and Odds Ratio Differences

• Likelihood ratios and odds ratios have different “reference groups”
  – LR: Reference group is the overall population; LR+ and LR- are used to obtain probabilities among those with positive and negative tests (dichotomous test, 2 LR)
  – OR: Reference group is the unexposed group; OR is used to obtain probability among those who are exposed (dichotomous exposure, 1 OR)

LR+ and OR

• LR+
  – Sensitivity = a / (a + c) = a / e
  – 1-Specificity = b / (b + d) = b / f
  – LR+ = Sensitivity / (1-Specificity) = (a / e) / (b / f)
    = (a * f) / (b * e)
  – ORExposed = (a * d) / (b * c)
• LR-
  – 1-Sensitivity = c / (a + c) = c / e
  – Specificity = d / (b + d) = d / f
  – LR- = (1-Sensitivity) / Specificity = (c / e) / (d / f)
    = (c * f) / (d * e)
  – ORUnexposed = (c * b) / (d * a)
LR and OR Summary

- Calculate 2 LR (LR+ and LR-)
  - Each is independent; knowing one provides no information about the other
- Usually calculate 1 OR (either OR for exposure or OR for lack of exposure)
  - OR for exposure and lack of exposure are reciprocals and knowing one means knowing the other

Formulas for Probability of Disease

- Both likelihood ratios and odds ratios can be used to estimate probability of disease

\[
\text{Likelihood Ratio: } \frac{LR \times p}{(LR \times p) + (1-p)}
\]
\[
\text{Odds Ratio: } \frac{OR \times p}{(OR \times p) + (1-p)}
\]

- The difference between odds ratio and likelihood ratio equations relates to p
  - P for LR: probability in population
  - P for OR: probability in reference group

LR and RR

- Relative risk for disease given a positive test
  - Risk given a positive test = \(\frac{a}{a+b} = \frac{a}{g}\)
  - Risk given a negative test = \(\frac{c}{c+d} = \frac{c}{h}\)
  - Relative risk = \(\frac{a}{g} \div \frac{c}{h} = \frac{(a \times h)}{(c \times g)}\)
- Primary differences between LR and RR
  - For likelihood ratios, work down columns of 2x2 table
  - For relative risks, work across rows
Why Do We Use LR Instead of OR / RR?

- LR vs OR: More efficient to be able to identify a pre-test probability and a set of test characteristics than it is to identify a probability of disease given a negative test (i.e., negative predictive value).
- LR vs RR:
  - RR more sensitive to pre-test probability than is LR
  - Arrow of causality generally different for RR and LR
    - We use RR for an exposure because exposure induces disease
    - We use an LR for a diagnostic test because for many tests, disease induces a change in some other biological marker that we then use to infer probability of disease.

4. Confidence Intervals for Sensitivity and Specificity

- Formulas for CI are available for means / proportions and for differences in means / proportions.
- Formulas also available for categorical and continuous variables.

<table>
<thead>
<tr>
<th>Mean/Proportion</th>
<th>Difference in Mean/Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>?</td>
</tr>
<tr>
<td>Continuous</td>
<td>?</td>
</tr>
</tbody>
</table>

- Which formula is appropriate for CI for sensitivity and specificity?

Formula for CI for Sensitivity/Specificity

- For sensitivity and specificity, use formula for confidence interval around a single proportion.
  - Proportion of positive tests given disease.
  - Proportion of negative tests given no disease.

- Common formula:

\[ S \pm z_{\alpha/2} \left( \frac{p(1-p)}{n} \right)^{0.5} \pm \frac{1}{2n} \]

where \( p \) = sensitivity or specificity, \( q = 1-p \); \( n \) = number of observations in which proportion was observed and \( z = \) standard Normal deviate associated with a 2-tailed probability \( \alpha \) (e.g., for 95% confidence, 1.96). NOTE, wider limit (i.e., \( \pm \frac{1}{2n} \)) always falls towards 0.5 (e.g., if proportion equals 0.6, wider limit stretches toward 0.5).
Formula for CI for Sensitivity/Specificity (2)

- Newcombe reviewed 7 formulae for calculating CI for a single proportion and reported that Wilson score confidence interval performs well

\[
\text{Wilson CL} = \frac{2np + z^2}{2(n + z^2)} \pm \frac{z}{2(n + z^2)} \sqrt{2npq}
\]

- where \( p \) = sensitivity or specificity; \( q = 1-p \); \( n \) = size of sample in which sensitivity or specificity was measured; and \( z \) = standard normal deviate associated with a 2-tailed probability \( \alpha \) (e.g., for 95% confidence, 1.96)

Example of Wilson CI for Sensitivity

- \( \beta_{1-42} \) pg/ml

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T^+ \leq 192 )</td>
<td>54</td>
<td>11</td>
</tr>
<tr>
<td>( T^- &gt; 192 )</td>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td>( n )</td>
<td>56</td>
<td>52</td>
</tr>
</tbody>
</table>

- Sensitivity = 54 / 56 = 0.9643

\[
\text{LL} = 2\frac{56 \cdot 0.9643 + 1.96^2 - 1.96 \sqrt{0.9643^2 + 4 \cdot 56 \cdot 0.9643 \cdot 0.0357}}{2 \cdot (56 + 1.96^2)} = 0.8788
\]

\[
\text{UL} = 2\frac{56 \cdot 0.9643 + 1.96^2 + 1.96 \sqrt{0.9643^2 + 4 \cdot 56 \cdot 0.9643 \cdot 0.0357}}{2 \cdot (56 + 1.96^2)} = 0.9902
\]

Example of Wilson CI for Specificity

- \( \beta_{1-42} \) pg/ml

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T^+ \leq 192 )</td>
<td>54</td>
<td>11</td>
</tr>
<tr>
<td>( T^- &gt; 192 )</td>
<td>2</td>
<td>41</td>
</tr>
<tr>
<td>( n )</td>
<td>56</td>
<td>52</td>
</tr>
</tbody>
</table>

- Specificity = 41 / 52 = 0.7885

\[
\text{LL} = 2\frac{52 \cdot 0.7885 + 1.96^2 - 1.96 \sqrt{0.7885^2 + 4 \cdot 52 \cdot 0.7885 \cdot 0.2115}}{2 \cdot (52 + 1.96^2)} = 0.6597
\]

\[
\text{UL} = 2\frac{52 \cdot 0.7885 + 1.96^2 + 1.96 \sqrt{0.7885^2 + 4 \cdot 52 \cdot 0.7885 \cdot 0.2115}}{2 \cdot (52 + 1.96^2)} = 0.8778
\]
Stata Commands

- Immediate form of Stata's confidence interval program (cii) calculates “conservative” (i.e., wider) Clopper Pearson confidence intervals
- The syntax is: cii [Total N] [N with pos/neg test]
  . cii 56 54

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>56</td>
<td>.9642857</td>
<td>.0247988</td>
<td>.876982 .9956452</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>.7884615</td>
<td>.0566348</td>
<td>.6529624 .8993855</td>
</tr>
</tbody>
</table>

Stata Commands, Wilson CI

. cii 56 54, wilson

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>56</td>
<td>.9642857</td>
<td>.0247988</td>
<td>.8788122 .9901506</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>.7884615</td>
<td>.0566348</td>
<td>.659679 .8775563</td>
</tr>
</tbody>
</table>

Stata Commands, diagti

- Similar intervals derivable by use of diagti (user written program that can be downloaded if you run: help diagt)
  . diagti 54 2 11 41

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>Pr(A)</th>
<th>52%</th>
<th>42%</th>
<th>61.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>Pr(+</td>
<td>A)</td>
<td>96.4%</td>
<td>87.7%</td>
</tr>
<tr>
<td>Specificity</td>
<td>Pr(-</td>
<td>N)</td>
<td>78.8%</td>
<td>65.3%</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>Pr(A</td>
<td>+)</td>
<td>83.1%</td>
<td>71.7%</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>Pr(N</td>
<td>-)</td>
<td>95.3%</td>
<td>84.2%</td>
</tr>
</tbody>
</table>

- Also reports “conservative” Clopper Pearson confidence intervals
Confidence Intervals for LR

• As already demonstrated, can use similar calculations to derive likelihood ratios and relative risks, but work down columns of 2x2 table for likelihood ratios and across rows for relative risks
• Can thus rearrange formula for approximate confidence intervals for relative risks (Rothman, p. 243-4) so it can be applied to likelihood ratios
• Relative risks and likelihood ratios are distributed log normal
  – Calculate their confidence interval by calculating interval for log of ratios and exponentiating log interval

Formulas for Ln LR and SE Ln LR

• The log of the likelihood ratio can be calculated directly by taking ln(LR), but using the notation in the 2x2 Table in slide 23, it also equals:
  \[ \ln LR^+ = \ln(a) + \ln(f) - (\ln(b) + \ln(e)) \]
  \[ \ln LR^- = \ln(c) + \ln(f) - (\ln(d) + \ln(e)) \]
• The standard error of the log of the likelihood ratio equals:
  \[ \text{SEln}(LR^+) = \sqrt{\frac{1}{a} + \frac{1}{b} - \left(\frac{1}{e} + \frac{1}{f}\right)} \]
  \[ \text{SEln}(LR^-) = \sqrt{\frac{1}{c} + \frac{1}{d} - \left(\frac{1}{e} + \frac{1}{f}\right)} \]

Formulas for LR CI

• Formula for lower and upper limits of log of likelihood ratio equals:
  \[ \text{Log(LR)} CI = \ln(LR) \pm (z \times \text{SEln}(LR)) \]
• Formula for lower and upper limits of the likelihood ratio equals:
  \[ LR CI = \exp^{\ln(LR)} \pm (z \times \text{SEln}(LR)) \]
Example, LR+ for Aβ_{1-42} pg/ml ≤ 192

\[
\begin{align*}
\text{LR}^+ &= \frac{(54 \times 52)}{(11 \times 56)} = 4.5584 \\
\text{Log (LR)}^+ &= \ln(54) + \ln(52) - \ln(11) - \ln(56) = 1.5169808 \\
\text{SE, Log(LR)}^+ &= \left(\frac{1}{54} + \frac{1}{11} - \frac{1}{56} - \frac{1}{52}\right)^{0.5} = 0.269604 \\
\text{LL, log(LR)}^+ &= \exp(0.96981842) = 2.6907 \\
\text{UL, LR}^+ &= \exp(2.0441432) = 7.7225
\end{align*}
\]

Example, LR- for Aβ_{1-42} pg/ml ≤ 192

\[
\begin{align*}
\text{LR}^- &= \frac{(2 \times 52)}{(41 \times 56)} = 0.04529617 \\
\text{Log (LR)}^- &= \ln(2) + \ln(52) - \ln(41) - \ln(56) = -3.0945329 \\
\text{SE, Log(LR)}^- &= \left(\frac{1}{2} + \frac{1}{41} - \frac{1}{56} - \frac{1}{52}\right)^{0.5} = 0.69807043 \\
\text{LL, log(LR)}^- &= -3.0945329 - (1.96 \times 0.69807043) = -4.4627509 \\
\text{UL, LR}^- &= \exp(-1.7263149) = 0.1779
\end{align*}
\]

Stata Command for CI for LR+

\[
\text{csi 54 11 2 41}
\]

\begin{tabular}{c|c|c|c}
\hline
 & Exposed & Unexposed & Total \\
\hline
Cases & 54 & 11 & 65 \\
Noncases & 2 & 41 & 43 \\
\hline
Total & 56 & 52 & 885 \\
\hline
Risk & .9642857 & .2115385 & .6018519 \\
Point Estimate & .7527473 & .6315702 & .8705077 \\
95\% CI & .5315702 & .8739243 & \\
Risk diff & 4.558442 & 2.690772 & 7.722464 \\
Risk ratio & .780268 & .6283594 & .805077 \\
Attr frac ex & .780268 & .6283594 & .805077 \\
Attr frac pop & .6485207 & & \\
\hline
\end{tabular}

\text{chi2(1) = 63.78 Pr=chisq = 0.0000}
5. Sample Size and Test Characteristics

- Formulas for sample size are available for means / proportions and for differences in means / proportions
- Formulas also available for categorical and continuous variables

<table>
<thead>
<tr>
<th>Formula</th>
<th>Mean/Proportion</th>
<th>Difference in Mean/Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Continuous</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

- Which formula is appropriate for estimating sample size for sensitivity and specificity?

Sample Size for Sensitivity and Specificity

- Calculating a sample size for a proportion; thus want to be in categorical row
- Depending on what we want to establish, might use formulas for either column of table
  - Might want to establish a maximum length of one of the half intervals (e.g., less than or equal to 0.05)
  - Might want to ensure that the resulting interval excludes some minimum value (e.g., expect the point estimate to be 0.9 and want to ensure that we can be 95% confident it is greater than 0.8)
  - Might want to ensure that the test has greater sensitivity (or specificity) than a second test
- In the following, we work through the first option (others are in reading)
Sample Size That Ensures That Wider of 2 Confidence Limits Is No Longer than a Specified Length

- Can estimate sample size by rearranging equation for Wilson score confidence limits and solving for \( N \). The resulting quadratic equation has following roots:
  
  \[
  a = 4p^2 + 4L^2 - 8pL \\
  b = 4pz^2 + 8L^2z^2 - (8pLz^2 + 4Lz^2 + 4pqz^2) \\
  c = 4L^2z^4 - 4Lz^4 
  \]

  where \( p = \) sensitivity or specificity; \( q = 1 - p \); \( L = p + \) the maximum length; and \( z = \) the sum of the standard Normal deviates associated with a 2-tailed \( \alpha \) (i.e., for 95% confidence, 1.96) and a 1-tailed \( \beta \) (i.e., for 80% power, 0.84)

Sample Size (2)

- Solving quadratic equation for \( n \), usually yields 2 estimates for the upper limit (point estimate + maximum length) and 2 for the lower limit (point estimate - maximum length)
- Usually, either 1 or 2 of these 4 estimates will be positive
  - If only 1 is positive, it represents the sample size. If 2 are positive, the larger represents the sample size

Sample Size Example, Sensitivity

- If we plan for a sensitivity of 0.9 and want the wider of the 2 confidence limits to be no longer than 0.05 (i.e., if we want the lower 95% limit for a sensitivity of 0.9 to be no smaller than 0.85 and the upper limit to be no larger than 0.95), the two positive roots for the quadratic equation would be 73 and 196
- The resulting sample size would be 196
- Table D1 reports sample sizes for selected proportions and selected maximum lengths of the confidence limits. For example, if we plan for a sensitivity of 0.75 and want limits that are no wider than 0.1, the sample size would be 88.
Sample Size Table *

<table>
<thead>
<tr>
<th>Target Proportion</th>
<th>Maximum Length of Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.025</td>
</tr>
<tr>
<td>0.60 or 0.40</td>
<td>1503</td>
</tr>
<tr>
<td>0.70 or 0.30</td>
<td>1349</td>
</tr>
<tr>
<td>0.80 or 0.20</td>
<td>1072</td>
</tr>
<tr>
<td>0.90 or 0.10</td>
<td>673</td>
</tr>
<tr>
<td>Max SS</td>
<td>1537</td>
</tr>
</tbody>
</table>

* Table entries represent sample sizes. "Max SS" represents the largest sample size required by any proportion that is greater than 0 and less than 1.

Only an Approximation

- Result is rough approximation
  - (Based on simulation) tends to understate needed sample size as proportion approaches 0.5 or tolerance approaches 0
- Stata’s "power oneprop" also seems to have problems

Sample Size and Cohort Designs

- Suppose we plan to evaluate the sensitivity and specificity of a test using a cohort design (i.e., we cannot determine ahead of time who is truly diseased and who is truly nondiseased)
- What are the implications for sample size?
  - Total number of patients in sample needs to be the larger of:
    \[ N_{\text{sens}} / \text{Prevalence} \text{ or } N_{\text{spec}} / (1-\text{Prevalence}) \]
Sample Size and Cohort Designs (2)

• Thus, if we select the maximum sample size for ± 0.05 confidence intervals around a sensitivity and specificity (n=385 with disease and 385 without disease), and if we expect a prevalence of disease of 0.2 in the population used to evaluate the test, we will need larger of:
  – Sensitivity: 385/0.2 = 1925
  – Specificity: 385/0.8 = 481
• With a sample of 1925, can expect to perform test in 385 people with disease and 1540 without disease

Sample Size for a Likelihood Ratio

• Can also take one of several approaches when calculating sample size for likelihood ratios
• Focus on a sample size for wider of 2 asymmetric limits to be no longer than some fixed length
• Rearranging equation for the likelihood ratio confidence limits and solving for N:

\[ n_s = \frac{z^2 \left( \frac{1}{p_1} + \frac{1}{p_0} - (1 + r) \right)}{\left( \ln \left( \frac{p_1}{p_0} \right) - \ln \left( \frac{p_2}{p_1} \pm m_l \right) \right)^2} \]

where \( m_l \) equals the maximum length of the confidence limit; and \( r \) equals the ratio of the number of those in whom disease is absent to the number of those in whom disease is present

Sample Size for a Test Characterized by Likelihood Ratios

• To determine sample size required for test, rather than for one of its likelihood ratios, compute sample size for each of expected likelihood ratios, and then use largest computed sample size
• For a sensitivity of 0.9 and a specificity of 0.45:
  – LR+ would be expected to be 1.6363; LR- would be expected to be 0.2222
• If want a confidence limit that is no longer than 0.2 around LR+ and no longer than 0.5 around LR- and if wanted 1 to 1 sampling of those with disease present and absent, the required sample size for LR+ would be 269, while the required sample size for LR- would be 29
6. Sankey and Henry Debate

- Evidence-based diagnostic test types routinely argue that healthcare professionals should become probabilists and use methods we’ve described for complex medical decision making.
- Sankey has always expressed doubt about this position, but I’ve tended to support it.
- I’ve been moving towards Sankey’s position for a number of reasons.

Lack of Evidence

- While those holding out this position tend to be evidence-based, there is no evidence that if physicians learned the material and used it in their daily practice, that they would be better physicians.
  - After lots of thought, I tend to see medical education as pattern recognition, not number crunching.

“Advanced” Technical Issues

- Simple version of math that we teach is appropriate for decision making when there are two options (e.g., disease or no disease).
  - Math is much more complicated when there are more than 2 options.
- For complicated decision making, the simple math tends to assume that the test characteristics are independent of one another.
  - Not much evidence exists on independence or lack thereof of different tests’ results.
- Not clear that physicians have a good sense of their treatment threshold.
Important For Researchers Developing or Evaluating Tests

- Fact that clinicians may be pattern recognizers and not probabilists does not imply we can ignore principles about what is and what is not a positive test
- When developing and analyzing tests, principles essential to determining when tests should be used and how they should be interpreted