Topics in Dichotomous Test Evaluation

Henry Glick

Epi 550

January 22, 2020

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

   ?? To obtain a precise estimate of probability of disease??

2) To target one of several treatments to a patient (precision medicine)
3) To understand disease process simply for sake of knowing
4) To avoid malpractice
5) 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 5 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)
- Development of prediction rules
- Verification bias
- Comparison of 2x2 and SSLR approaches
- Graphing results of optimal test selection
- Choice among tests

Outline for Today

- Treatment threshold and difference in outcomes
- 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 Difference in Outcomes

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

No test- No treat	Test and Treat if Test result is positive	Treat

0 Probability of disease 1

^ ^
TT TTT

Logic Behind Thresholds

- Value of not testing/not treating highest as pretest probability of disease approaches 0 and lowest as it approaches 1
- Value of treating empirically highest/lowest as pretest probabilities of disease approach 1/0
- Value of testing/treating highest in the middle of it's range and decreases as pretest probabilities approach 0 and 1
- No Test/Test Threshold (TTT) defined as probability where expected value of not testing/not treating equals expected value of testing/treating
- Test/Treat Theshold (TT) defined as probability where expected value of testing/treating equals expected value of treating empirically

Other Thresholds

- Other thresholds exist
  - e.g., test and treat; withdraw treatment if test result is negative
    - Used when expected value of testing, treating, and withdrawing treatment after a negative test is greater than expected value 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℃
- 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 OF DISEASE

Likelihood (Probability) of Disease

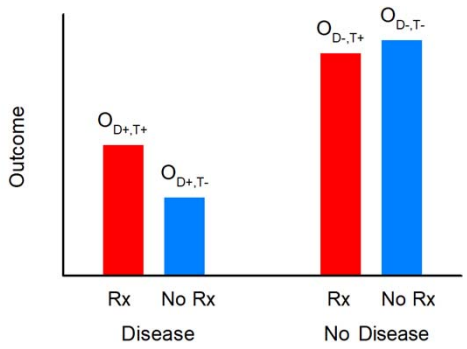
- All else equal, empiric therapy more likely indicated for higher likelihoods (probabilities) of disease
- Empiric therapy less likely indicated for lower likelihoods of disease

DIFFERENCE IN VALUE OF OUTCOMES

Difference in Value of Outcomes

- All else equal, when difference in value between correctly treating vs incorrectly withholding treatment increases compared to difference between correctly withholding treatment vs incorrectly treating, more likely that empirical treatment indicated
- When difference in value between correctly treating vs incorrectly withholding treatment decreases compared to difference between correctly withholding treatment vs incorrectly treating, more likely that withholding treatment indicated

Why ***DIFFERENCE*** in Outcomes?



Treatment Threshold

- Can combine information about probability of disease and difference in outcomes to identify a probability of disease (p*) where expected outcomes from treating and withholding treatment are equal
  - If probability of disease is above p*, expected outcome from treatment exceeds expected outcome from withholding treatment, and treatment is indicated
  - If below p*, reverse is true and withholding treatment indicated

Definitions

$O_{D+,Rx+} = O_{TP}$ = Value of outcome given treatment when disease is present

$O_{D+,Rx-} = O_{FN}$ = Value of outcome given withholding of treatment when disease present

$O_{D-,Rx-} = O_{TN}$ = Value of outcome given withholding of treatment when disease is absent

$O_{D-,Rx+} = O_{FP}$ = Value of outcome given treatment when disease is absent

$O_{TP} - O_{FN} = \Delta O_{D+}$ = Difference in outcome | disease

$O_{TN} - O_{FP} = \Delta O_{D-}$ = Difference in outcome | no disease

ΔO_{D-} and ΔO_{D+}

- ΔO_{D-} sometimes (pessimistically?) referred to as "(incremental) cost of false positive (C_{fp})"
  - What is lost when we incorrectly treat
- Equally justifiable to refer to it as "(incremental) benefit of true negative (B_{tn})"
  - What is gained when we correctly withhold treatment
- ΔO_{D+} sometimes (pessimistically?) referred to as "(incremental) cost of false negative (C_{fn})"
  - What is lost when we incorrectly withhold treatment
- Equally justifiable to refer to it as "(incremental) benefit of true positive (B_{tp})"
  - What is gained when we correctly treat
- Same quantities independent of optimism/pessimism

Treatment Threshold (p*)

- Set expected outcome of treatment equal to expected outcome of no treatment

Expected outcome of treatment (EO_{Treat})

$$EO_{Treat} = pO_{TP} + (1-p)O_{FP}$$

Expected outcome of no treatment ($EO_{NoTreat}$)

$$EO_{NoTreat} = pO_{FN} + (1-p)O_{TN}$$

Treatment threshold = $p^* = (EO_{Treat} = EO_{NoTreat})$

Deriving Treatment Threshold

Solve for p^* such that $(EO_{Treat} = EO_{NoTreat})$

[1] $pO_{TP} + (1-p)O_{FP} = pO_{FN} + (1-p)O_{TN}$

[2] $pO_{TP} + O_{FP} - pO_{FP} = pO_{FN} + O_{TN} - pO_{TN}$

[3] $(pO_{TP} - pO_{FN}) + (pO_{TN} - pO_{FP}) = (O_{TN} - O_{FP})$

[4] $p[(O_{TP} - O_{FN}) + (O_{TN} - O_{FP})] = (O_{TN} - O_{FP})$

[5] $p\,(\Delta O_{D+} + \Delta O_{D-}) = \Delta O_{D-}$

[6] $p^* = \Delta O_{D-} / (\Delta O_{D+} + \Delta O_{D-})$

Treatment Decisions and Threshold

$\Delta O_{D-} / (\Delta O_{D+} + \Delta O_{D-}) =$ 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 outcome from withholding treatment exceeds that from treatment
  - If probability is greater than threshold, treat empirically because expected outcome from treatment exceeds that from withholding 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 yields posttest probabilities that are greater/less than pretest probabilities:
    - Negative tests should yield posttest probabilities that are below treatment threshold
    - Positive tests should yield posttest probabilities that exceed treatment threshold

Definition of Difference in Outcomes

- 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)
  - E.g., a false positive costs an extra $100 as compared to a true negative (a true negative saves $100 as compared to a false positive)
- Also can be estimated by use of a cost-effectiveness (NMB) framework (separate calculation of incremental costs (c) and outcomes (e)):
  - $\Delta O_{D+} = (W\ e_{D+}) - c_{D+}$
  - $\Delta O_{D-} = (W\ e_{D-}) - c_{D-}$

    where W = willingness to pay

Cost-Effectiveness Equivalent

- Under a cost-effectiveness framework (in which ΔO_{D+} equals a combination of c_{D+} and e_{D+} and ΔO_{D-} equals a combination of c_{D-} and e_{D-}), and W equals maximum willingness to pay:

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

Ratio of Differences Sufficient

- Many people uncomfortable with identifying absolute difference in outcomes among those with and without disease
- Good news: [If we can ignore the cost of the test] ratio of difference in outcomes more important than absolute magnitudes
  - e.g., when defining treatment threshold, difference in outcomes among those without disease thought to be 1/3 the difference in outcomes among those with disease, know that:

$$\Delta O_{D-} = 1/3\ \Delta O_{D+}$$

➔ $p^* = 1/3\ \Delta O_{D+} / ((1/3\ \Delta O_{D+}) + \Delta O_{D+}) = 0.25$

Treatment Threshold and Differences in Outcomes

- If instead have an idea of treatment threshold, can infer relative valuation of ΔO_{D-} and ΔO_{D+}
  - If our $p^* = 0.25$, then:

$$0.25 = \Delta O_{D-} / (\Delta O_{D+} + \Delta O_{D-})$$
$$0.25\Delta O_{D+} + 0.25\Delta O_{D-} = \Delta O_{D-}$$
$$0.25\Delta O_{D+} = 0.75\ \Delta O_{D-}$$
$$1/3\ \Delta O_{D+} = \Delta O_{D-}$$

- Implication: Implicitly making assumptions about ΔO_{D-} and ΔO_{D+} whenever a treatment decision is made

TT and TTT

- Defining the do nothing / test threshold (TT) and test / treat threshold (TTT) requires information about the test that is not required for identifying p*
  - Test characteristics
  - Cost of test
- Formulas for deriving TT and TTT equal:

$$TT = p_{TT} = \frac{(1\text{-spec})\,\Delta O_{D-} + TC}{\text{sens}\,\Delta O_{D+} + (1\text{-spec})\,\Delta O_{D-}}$$

$$TTT = p_{TTT} = \frac{\text{spec}\,\Delta O_{D-} - TC}{(1\text{-sens})\,\Delta O_{D+} + \text{spec}\,\Delta O_{D-}}$$

* where TC = cost of test

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 differences in result from using sensitivity and specificity vs likelihood ratios for positive and negative tests

  ➔ Corollary: when test result is continuous, post-test probabilities from likelihood ratios can differ from those 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

	Disease		
	Present	Absent	Total
Test result positive	(a) True positive (TP)	(b) False positive (FP)	Test + (a + b) = g
Test result Negative	(c) False negative (FN)	(d) True negative (TN)	Test - (c + d)=h
Total	All with disease (a + c) = e	All normals (b + d) = f	

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 (or exposed) group
    - OR is used to obtain probability among those who are exposed (unexposed) (dichotomous exposure, 1 OR)

LR+ and OR

- LR+: Sensitivity / (1-Specificity)
  - Sensitivity = a / (a + c) = a / e
  - 1-Specificity = b / (b + d) = b / f
  - LR+ = (a / e) / (b / f)
    = (a * f) / (b * e)
  - $OR_{Exposed}$ = (a * d) / (b * c)
- LR-: 1-Sensitivity / Specificity
  - 1-Sensitivity = c / (a + c) = c / e
  - Specificity = d / (b + d) = d / f
  - LR- = (c / e) / (c / f)
    = (c * f) / (d * e)
  - $OR_{Unxposed}$ = (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 Calculating Probability of Disease

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

Likelihood Ratio	Odds Ratio
$\frac{LR \times p}{(LR \times p) + (1-p)}$	$\frac{OR \times p}{(OR \times p) + (1-p)}$

- Difference between likelihood ratio and odds ratio equations relate to LR/OR AND 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 = a / (a + b) = a / g
  - Risk given a negative test = c / (c + d) = c / h
  - Relative risk = (a / g) / (c / h) = (a * h) / (c * 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., 1-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

	Mean/Proportion	Difference in Mean/Proportion
Categorical	?	?
Continuous	?	?

- In which cell does formula for CI for sensitivity and specificity fall?

Example of Common Formula CI for Sensitivity

- Auditory biological marker *

	D+	D-
T+	18	1
T-	2	19
	20	20

- Sensitivity = 18 / 20 = 0.90
- Specificity = 19/20 = 0.95

* Kraus N, et al. Auditory biological marker of concussion in children. Scientific Reports. 6:39009 | DOI: 10.1038/srep39009

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 20 18
```

Variable	Obs	Mean	Std. Err	[95% Conf. Interval]	
	20	.9	.067082	.6830173	.9876515

```
. cii 20 19
```

Variable	Obs	Mean	Std. Err	[95% Conf. Interval]	
	20	.95	.048734	.7512672	.9987349

Formula for CI for Sensitivity/Specificity (2)

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

$$\text{Wilson CL} = \frac{2np + z^2 \pm z\sqrt{z^2 + 4npq}}{2\,(n + z^2)}$$

- 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 α (e.g., for 95% confidence, 1.96)

Example of Wilson CI for Sensitivity

- Auditory biological marker

	D+	D-
T+	18	1
T-	2	19
	20	20

- Sensitivity = 18 / 20 = 0.90

$$\text{LL} = \frac{2*20*0.9 + 1.96^2 - 1.96\sqrt{1.96^2 + 4*20*0.9*0.1}}{2\,(20 + 1.96^2)} = 0.6990$$

$$\text{UL} = \frac{2*20*0.9 + 1.96^2 + 1.96\sqrt{1.96^2 + 4*20*0.9*0.1}}{2\,(20 + 1.96^2)} = 0.972$$

Stata Commands, Wilson CI

. cii 20 18,wilson

Variable	Obs	Mean	Std. Err	[95% Conf.	Interval]
	20	.9	.067082	.6989664	.9721335

VERSUS Clopper Pearson: 0.683 – 0.988

Example of Wilson CI for Specificity

- Auditory biological marker

	D+	D-
T+	18	1
T-	2	19
	20	20

- Specificity = 19 / 20 = 0.95

$$LL = \frac{2*20*0.95 + 1.96^2 - 1.96\sqrt{1.96^2 + 4*20*0.95*0.05}}{2\,(20 + 1.96^2)} = 0.7639$$

$$UL = \frac{2*20*0.95 + 1.96^2 + 1.96\sqrt{1.96^2 + 4*20*0.95*0.05}}{2\,(20 + 1.96^2)} = 0.9911$$

Stata Commands, Wilson CI

. cii 20 19,wilson

Variable	Obs	Mean	Std. Err	[95% Conf.	Interval]
	20	.95	.048734	.7638688	.9911186

- Versus Clopper-Pearson: .751 to 0.999

Stata Commands, diagti

- Use of diagti also yields clopper pearson intervals
  - Diagti a user written program that can be downloaded if you run: help diagt

. diagti 18 2 1 19

Prevalence	Pr(A)	50%	34%	66.2%
Sensitivity	Pr(+| A)	90%	**68.3%**	**98.8%**
Specificity	Pr(-| N)	95%	**75.1%**	**99.9%**
Positive predictive value	Pr(A| +)	94.7%	74%	99.9%
Negative predictive value	Pr(N| -)	90.5%	69.6%	98.8%

- 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

Stata Command for CI for LR+

csi N_{tp} N_{fp} N_{fn} N_{tn}

csi 18 1 2 19

	Exposed	Unexp	Total	
Cases	18	1	19	
Noncases	2	19	21	
Total	20	20	40	
Risk	.9	.05	.475	
	Point Estimate		95% CI	
Risk diff	.85		.6874884	1.012512
Risk ratio	18		**2.649729**	**122.2766**
Attr frac ex	.9444444		.6226603	.9918218
Attr frac pop	.8947368			

chi2(1) = 28.97 Pr>chi2 = 0.0000

Exposed = D+; Cases = T+ Unexposed = D-; Noncases = T-

Stata Command for CI for LR-

csi N_{fn} N_{tn} N_{tp} N_{fp}

csi 2 19 18 1

	Exposed	Unexp	Total	
Cases	2	19	21	
Noncases	18	1	19	
Total	20	20	40	
Risk	.1	.95	.525	
	Point Estimate		95% CI	
Risk diff	-.85		-1.012512	-.6874884
Risk ratio	**.1052632**		**.0281583**	**.393502**
Prev frac ex	.8947368		.606498	.9718417
Prev frac pop	.4473684			

chi2(1) = 28.97 Pr>chi2 = 0.0000

Exposed = D+; Cases = T+ Unexposed = D-; Noncases = T-

Stata Commands, diagti

- Obtain similar intervals using diagti

Prevalence	Pr(A)	50%	34%	66.2%
Likelihood ratio (+)	Pr(+|A)/(Pr(+N)	18	2**.65%**	**122%**
Likelihood ratio (-)	Pr(-|A)/PR(-|N)	.105	**.0282**	**.394**
Odds ratio	LR(+)/LR(-)	171	16.7	
Positive predictive value	Pr(A|+)	94.7%	74%	99.9%
Negative predictive value	Pr(N|-)	90.5%	69.6%	98.8%

Distributions of Relative Risks and 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 29, 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:

  $SE\ln(LR+) = ((1/a) + (1/b) - ((1/e) + (1/f)))^{0.5}$

  $SE\ln(LR-) = ((1/c) + (1/d) - ((1/e) + (1/f)))^{0.5}$

Formulas for LR CI

- Formula for lower and upper limits of log of likelihood ratio equals:

  $\text{Log(LR) CI} = \ln(LR) \pm (z * SE_{\ln(LR)})$

- Formula for lower and upper limits of the likelihood ratio equals:

  $\text{LR CI} = \exp^{\ln(LR) \pm (z * SE\ln(LR))}$

Example, LR+ for Auditory Biological Marker

$LR+ = (18*20) / (1 * 20) = 18.0$

$\text{Log (LR+)} = \ln(18) + \ln(20) - \ln(1) - \ln(20) = 2.8903718$

$\text{SE, Log(LR+)} = (1/18 + 1/1 - (1/20 + 1/20))^{0.5} = 0.97752522$

$\text{LL, log(LR+)} = 2.8903718 - (1.96 * 0.97752522) = 0.974445753$

$\text{UL, log(LR+)} = 2.8903718 + (1.96 * 0.97752522) = 4.806286$

$\text{LL, LR+} = \exp(0.974445753) = 2.6497294$

$\text{UL, LR+} = \exp(4.806286) = 122.27664$

Example, LR- for Auditory Biological Marker

LR- = (2*20) / (19 * 20) = 0.10526316

Log (LR-) = ln(2) + ln(20) - ln(19) - ln(20) = -2.2512918

SE, Log(LR-) = $(1/2 + 1/19 - (1/20 + 1/20))^{0.5}$ = 0.672779

LL, log(LR-) = -2.2512918 - (1.96 * 0.672779) = -3.5699144

UL, log(LR-) = -2.2512918 + (1.96 * 0.672779) = -0.93266919

LL, LR- = exp(-3.5699144) = 0.02815826

UL, LR- = exp(-0.93266919) = 0.39350197

Henry's Concerns, Audio Biological Marker

- (Sample size was determined based on ability to detect group differences in F_0[12], not variability around sensitivity and specificity (results being touted in article)
  - Uncertainty assessment suggests substantially more evaluation is needed before we can be confident of test's operating characteristics
- Controls were a community sample with no reported history of brain injury
  - To avoid potential spectrum bias, test should be evaluated in population in which it will be used
    - Cases: children with head injuries who, upon workup, we find had a concussion; controls: children with head injuries who, upon workup, we find did not have concussion

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

	Mean/Proportion	Difference in Mean/Proportion
Categorical	?	?
Continuous	?	?

- 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 $\pm$ the maximum length; and z = the sum of the standard Normal deviates associated with a 2-tailed α (i.e., for 95% confidence, 1.96) and a 1-tailed β (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 *

Target Proportion	Maximum Length of Confidence Limits		
	0.025	0.05	0.10
0.60 or 0.40	1503	381	97
0.70 or 0.30	1349	350	93
0.80 or 0.20	1072	289	81
0.90 or 0.10	673	196	62
Max SS	1537	385	97

* 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" (power analysis for a one-sample proportion test) 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_{Sens} / Prevalence or N_{Spec} / 1-Prevalence

Sample Size and Cohort Designs (2)

- Thus, if we select the maximum sample size for $\pm$ 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, 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 ito be no longer than some fixed length
- Rearranging equation for the likelihood ratio confidence limits and solving for N:

$$n_d = \frac{z^2 \left(\frac{1}{p_d} + \frac{1}{r p_h} - \left(1 + \frac{1}{r}\right)\right)}{\left(\ln\left(\frac{p_d}{p_h}\right) - \ln\left(\frac{p_d}{p_h} \pm ml\right)\right)^2}$$

$$n_h = r * n_d$$

- where ml 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 clinicians learned the material and used it in their daily practice, that they would be better clinicians
  - After lots of thought, I tend to see medical education as pattern recognition, not number crunching
    - Because some excellent clinicians are NOT excellent at math

"Advanced" Technical Issues

- Simple version of math 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
- Simple math tends to assume that test characteristics are independent of one another
  - Not much evidence exists on independence or lack thereof of different test's results
- Not clear that clinicians 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