Evaluation of diagnostic agents: Sample Size Estimation Using Exact Methods

Stefano Vezzoli, CROS NT
Agenda

- Evaluation of the performance of a dichotomous diagnostic agent
- Sample size calculation for a non-inferiority trial
- Normal-approximate approach
- Exact approach

Questions
Dichotomous diagnostic tests

Two possible results and two possible disease states. We can tabulate them:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP (True Positives)</td>
<td>FP (False Positives)</td>
</tr>
<tr>
<td>Negative</td>
<td>FN (False Negatives)</td>
<td>TN (True Negatives)</td>
</tr>
</tbody>
</table>

**Sensitivity**: how good a test is at correctly identifying subjects who have the disease

\[
\text{Sensitivity} = \frac{TP}{(TP+FN)}
\]

**Specificity**: how good the test is at correctly identifying subjects without the disease

\[
\text{Specificity} = \frac{TN}{(TN+FP)}
\]
Dichotomous diagnostic tests

Example

Sensitivity and Specificity of the Semiquantitative Latex Agglutination D-Dimer Assay for the Diagnosis of Acute Pulmonary Embolism as Defined by Computed Tomographic Angiography.


- “Pulmonary angiography is the diagnostic reference standard, but it is invasive and associated with some morbidity and mortality. An alternative approach for patients with nondiagnostic lung scans is noninvasive testing for deep venous thrombosis”.

- “A positive test was defined as a D-dimer level greater than 250 ng/mL”. 
Dichotomous diagnostic tests

Example

Table 1. Diagnosis of Acute Pulmonary Embolism*

<table>
<thead>
<tr>
<th>D-dimer results</th>
<th>CT angiography results</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>142</td>
<td>470</td>
</tr>
<tr>
<td>Negative</td>
<td>30</td>
<td>304</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>774</td>
</tr>
</tbody>
</table>

*CT = computed tomographic.

- Sensitivity = TP / (TP+FN) = 142 / 172 = 0.83
- Specificity = TN / (TN+FP) = 304 / 774 = 0.39
Regulatory considerations

Points to Consider on the Evaluation of Diagnostic Agents (CPMP, 2001)

Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests (FDA, 2007)

- Sensitivity
- Specificity

Sensitivity AND Specificity are designated as appropriate primary endpoints for the evaluation of diagnostic performance of an experimental agent.
Regulatory considerations

Example

Requirements established by FDA for a approval of a rapid HIV test:

- **Sensitivity**: 98% (lower bound of the 95% confidence interval)

- **Specificity**: 98% (lower bound of the 95% confidence interval)
**Example of study**

- **Non-inferiority** trial aiming to show acceptable levels of inferiority when comparing the diagnostic agent with an **absolute standard reference**

  The absolute standard by definition can truly reflect the presence or absence of the target disease (sens = 1, spec = 1)

- Minimally acceptable Sensitivity \((\text{sens}_0)\) and Specificity \((\text{spec}_0)\) should be specified in advance

- The study will test the one-sided inferiority hypothesis

  \[ H_0 : \{ \text{sens} \leq \text{sens}_0 \text{ or spec} \leq \text{spec}_0 \} \]
The hypothesis can be tested by calculating a joint rectangular confidence region for (sens, spec). Cross-product of two one-sided confidence intervals.
Hypothesis testing
Hypothesis testing

- **Adjustment for multiplicity** resulting from the assessment of two primary endpoints (sensitivity and specificity) is not required because both primary hypotheses need to be rejected.
- This procedure **inflates the type II error** (here: falsely accepting that at least one hypothesis is true). This inflation must be taken into account for a proper estimation of the sample size for the trial.

Data from diseased and non-diseased subjects, on which Sensitivity and Specificity are respectively estimated, are independent
Hypothesis testing

Sensitivity and Specificity are estimated *separately*

If the power of each single test is $1-\beta^* = \sqrt{(1-\beta)}$, then the global power $1-\beta$ is ensured

Joint test (sens, spec):
- $\alpha = 0.05$
- $1-\beta = 0.8$

Each single test:
- $\alpha^* = 0.05$
- $1-\beta^* \approx 0.9$

From now on, we will consider the problem of sample size estimation for a single proportion
Normal-approximate approach

- Based on asymptotic theory
- Sample size can be calculated by using PROC POWER

```
proc power;
   onesamplefreq test=z method=normal
      sides = u
      alpha = 0.05
      nullproportion = 0.75
      proportion = 0.9
      ntotal = .
      power = 0.9;
run;
```

Under the normal approximation, 54 subjects are needed
If we compute the significance level of the normal-approximate test under the true binomial distribution...

```proc power;
    onesamplefreq test=z method=exact
        sides       = u
        alpha       = 0.05
        nullproportion = 0.75
        proportion  = 0.9
        ntotal      = 54
        power       = .;
run;
```

**\( \alpha^* = 0.0525 > 0.05 \)**
One may think that increasing the sample size will result in lowering the significance level. Unfortunately, the solution is not so obvious.
Normal-approximate approach

Using the **ODS** capabilities of PROC POWER, we can plot

- the actual Type I error rate
- the achieved power

as functions of the sample size

```
proc power;
  ods output plotcontent=PlotData;
  onesamplefreq test=z method=exact
    sides = u
    alpha = 0.05
    nullproportion = 0.75
    proportion = 0.9
    ntotal = 50
    power = .;
  plot x=n min=50 max=100 step=1;
run;
```
The Type I error rate fluctuates around the nominal level.
Normal-approximate approach

The power curve does not monotonically increase with increasing sample size

- $n = 54 \rightarrow 1-\beta^* > 0.9$
- $n = 56 \rightarrow 1-\beta^* < 0.9$
- $n = 58 \rightarrow 1-\beta^* > 0.9$
Normal-approximate approach

• The “saw-toothed” behavior of these functions is due to the discrete nature of the underlying binomial distribution
• Under the normal approximation, we must pay attention to the choice of a sample size that ensures adequate power and significance level
• This problem can be partly overcome using a test based on the true binomial distribution
  
  \[ \text{“exact” method} \]

• The nominal value \( \alpha^* \) is the upper bound for the true significance level
• We must only take care of power requirements
Exact approach

With PROC POWER the sample size is not available as result parameter under the exact approach (Clopper-Pearson CI)

We can still use PROC POWER for calculating the actual type I error rate and the achieved power as functions of $n$

```r
proc power;
  ods output plotcontent=PlotData;
  oneproportion freq test=exact method=exact
    sides = u
    alpha = 0.05
    nullproportion = 0.75
    proportion = 0.9
    ntotal = 50
    power = .;
  plot x=n min=50 max=100 step=1;
run;
```
The Exact approach

- Normal approximation
- Exact approach

The nominal significance level $\alpha^*$ is ensured

- $n = 54$, $H_0$ rejected if $x \geq 46 \rightarrow \alpha^* > 0.05$
- $n = 54$, $H_0$ rejected if $x \geq 47 \rightarrow \alpha^* < 0.05$
We can focus just on the power.

The sample size can be defined in two ways:

- the minimum $n$ which satisfies $1-\beta_A \geq 1-\beta^*$
- the minimum $n$ which satisfies $1-\beta_A \geq 1-\beta^*$, and the condition is also satisfied for any sample size larger than $n$
The %DSS macro

%DSS(alpha=,power=,sens=,min_sens=,spec=,min_spec=,cond=);

- **alpha**: required significance level $\alpha$
- **power**: required power $1-\beta$
- **sens**: expected sensitivity ($sens_1$)
- **spec**: expected specificity ($spec_1$)
- **min_sens**: minimally acceptable sensitivity ($sens_0$)
- **min_spec**: minimally acceptable specificity ($spec_0$)
- **cond**: sample size calculated under the weak condition (=W), the strong condition (=S), or both conditions (not specified)

Different values can be entered for all the numerical parameters
The %DSS macro

- $\alpha = 0.1$
- $1-\beta = 0.8$
- $\text{sens}_1 = 0.9$
- $\text{sens}_0 = 0.75$
- $\text{spec}_1 = 0.95$
- $\text{spec}_0 = 0.8$
- under both the conditions

vs 54 subjects needed under the normal approximation
ANY QUESTIONS?