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:

		True disease state: Present	True disease state: Absent
Test result	**Positive**	**TP** (True Positives)	**FP** (False Positives)
	Negative	**FN** (False Negatives)	**TN** (True Negatives)

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

TP / (TP+FN)

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

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.

Froehling D.A. et al. (2004) *Mayo Clinic Proceedings*; 79: 164-168.

- "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***

D-dimer results	CT angiography results		
	Positive	Negative	Total
Positive	142	470	612
Negative	30	304	334
Total	172	774	946

*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 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)

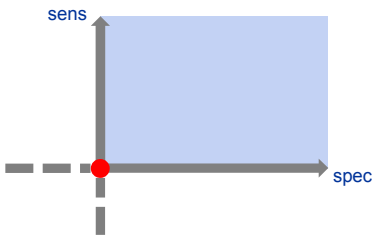
- Minimally acceptable Sensitivity ($sens_0$) and Specificity ($spec_0$) should be specified in advance
- The study will test the one-sided inferiority hypothesis

$$H_0 : \{ sens \leq sens_0 \text{ or } spec \leq spec_0 \}$$

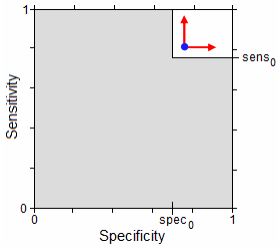
Hypothesis testing

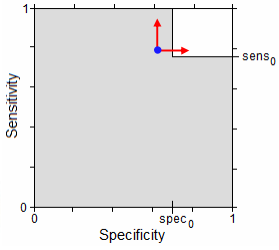
The hypothesis can be tested by calculating a joint **rectangular confidence region** for (sens, spec)

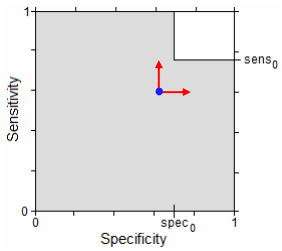
Cross-product of two one sided confidence intervals



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**

n_D n_{ND} calculated **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

Normal-approximate approach

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;
```

$\alpha^* = 0.0525 > 0.05$

Normal-approximate approach

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

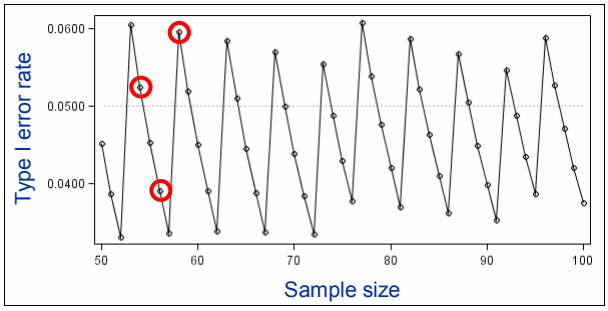
Using the **ODS** capabilities of PROC POWER, we can plot

- 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;
```

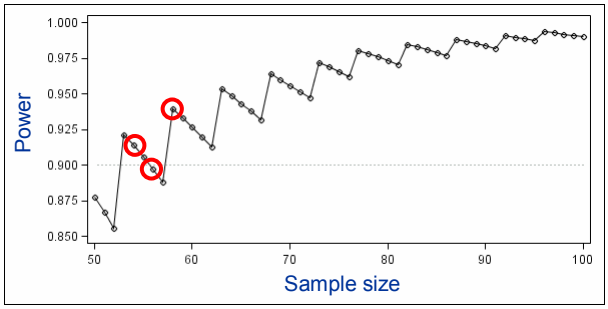
Normal-approximate approach



- $n = 54 \rightarrow \alpha^* > 0.05$
- $n = 56 \rightarrow \alpha^* < 0.05$
- $n = 58 \rightarrow \alpha^* > 0.05$

The Type I error rate fluctuates around the nominal level

Normal-approximate approach



- $n = 54 \rightarrow 1\text{-}\beta^* > 0.9$
- $n = 56 \rightarrow 1\text{-}\beta^* < 0.9$
- $n = 58 \rightarrow 1\text{-}\beta^* > 0.9$

The power curve does not monotonically increase with increasing sample size

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**

→ **“exact”** method

- **The nominal value α* 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)

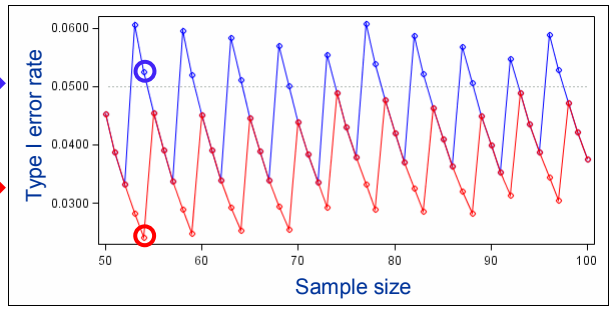
We can still use PROC POWER for calculating the actual type I error rate and the achieved power as functions of *n*

```
proc power;
  ods output plotcontent=PlotData;
  onesamplefreq 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;
```

Exact approach

Normal approximation

Exact approach



- $n = 54$, H_0 rejected if $x \geq 46 \rightarrow \alpha^* > 0.05$
- $n = 54$, H_0 rejected if $x \geq 47 \rightarrow \alpha^* < 0.05$

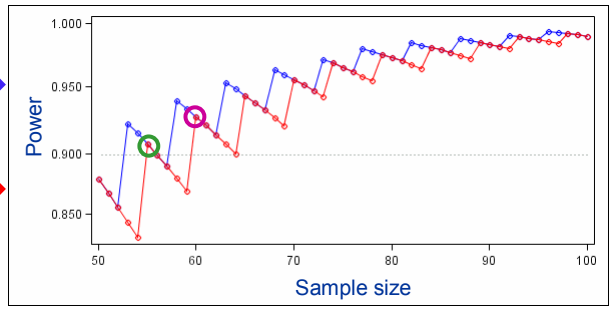
The nominal significance level α* is ensured

www.cros.it

Exact approach

Normal approximation

Exact approach



We can focus just on the power.

The sample size can be defined in two ways:

- the minimum n which satisfies $1\text{-}\beta_A \geq 1\text{-}\beta^*$
- the minimum n which satisfies $1\text{-}\beta_A \geq 1\text{-}\beta^*$, and the condition is also satisfied for any sample size larger than n

www.cros.it

The %DSS macro

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

- **alpha**: required significance level α
- **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

- $\alpha = 0.1$
- $1-\beta = 0.8$
- $sens_1 = 0.9$
- $sens_0 = 0.75$
- $spec_1 = 0.95$
- $spec_0 = 0.8$
- under both the conditions

Scenario	ALPHA Nominal	ALPHA Actual	POWER Nominal	POWER Actual	SENS Min	SENS Exp	SPEC Min	SPEC Exp	DIS N	DIS Crit	NOT DIS- N	NOT DIS- Crit
1 W	0.05	0.045	0.8	0.845	0.75	0.9	0.8	0.95	55	47	44	40
S	0.05	0.045	0.8	0.864	0.75	0.9	0.8	0.95	60	51	44	40

vs **54** subjects needed under the normal approximation

The end

ANY

QUESTIONS ?