

Power calculation for non-inferiority trials comparing two Poisson distributions

Corinna Miede, Accovion GmbH, Marburg, Germany
Jochen Mueller-Cohrs, Accovion GmbH, Marburg, Germany

Abstract

Rare events are often described by a Poisson distribution. In clinical research the examination of such rare events could be the basis of a non-inferiority trial. In order to plan such a trial the power of a statistical test comparing two Poisson distributions is required. The purpose of this paper is to present a method for calculating size and power of three statistical tests. The method can be easily realized with a short SAS[®] program. The paper will depict the approach graphically and theoretically.

1 Introduction

In the ICH E9 guideline "Statistical Principles for Clinical Trials" [1] the non-inferiority trial is described as a possible type of comparison in clinical research. Especially if a standard treatment still exists and a placebo controlled trial is not practicable because of ethical reasons the non-inferiority trial is an appropriate method [2]. Advantages, disadvantages and statistical details of non-inferiority trials are also described in [2]. In order to plan a non-inferiority trial the calculation of the power of a statistical test is required. For the comparison of two Poisson distributions several tests are available. This paper focuses on the following three tests:

- the likelihood ratio test,
- the score test,
- the exact conditional test.

The first two tests are based on asymptotic properties of the likelihood function. Therefore, in addition to the power of the tests the actual type I error rate is also of interest in applications with finite samples. In the following it is shown how the operating characteristic can be calculated exactly by summing the probability distribution function over the critical region. A realization with SAS is outlined. Finally the three tests are compared regarding size and power in a practical example.

2 Assumptions

Suppose a new treatment is to be compared with a control treatment in a parallel group study with n_1 individuals in the control treatment group and n_2 individuals in the new treatment group. The target variable observed on each individual is the number of occurrences of a certain event. The aim of the study is to demonstrate that the new treatment is not inferior or is superior to the control treatment in reducing the number of events. It is assumed that the number of events in each individual follows a Poisson process with mean μ_1 for the control group and mean μ_2 for the new treatment group. The mean values of the number of events refer to a certain time unit, e.g. a year. If the observation time of the j th individual in group i is t_{ij} then the total number of events Y_i in group i follows a Poisson distribution with mean

$$\lambda_i = m_i \mu_i \quad i = 1, 2 \quad \text{where} \quad m_i = \left(\sum_{j=1}^{n_i} t_{ij} \right).$$

m_i is the total observation time in group i . Particularly, if all individuals are observed over a unit time interval then m_i is equal to the sample size n_i in group i . The probability distribution function for the total number of events in group i is thus

$$\Pr(Y_i = y) = \frac{\exp(-\lambda_i) \lambda_i^y}{y!} \quad i = 1, 2.$$

The following one-sided null hypothesis H_0 is to be tested against the alternative H_1 :

$$H_0 : \mu_2/\mu_1 \geq \rho \quad \text{versus} \quad H_1 : \mu_2/\mu_1 < \rho.$$

If the ratio ρ is equal to or less than 1.0 the objective of the test is to show superiority of the new treatment. If ρ is greater than 1.0 the objective is to show non-inferiority of the new treatment with respect to the non-inferiority margin ρ .

3 Theoretical background

Let y_i denote the observed total number of events in group i . Further let

$$y_0 = y_1 + y_2,$$

$$\gamma = (m_2/m_1) \rho.$$

The likelihood-ratio statistic is

$$G^2 = 2 [y_1 \ln y_1 + y_2 \ln(y_2/\gamma) - y_0 \ln(y_0/(1 + \gamma))]$$

where $\ln$ denotes the natural logarithm and $y \ln y$ is defined to be zero if y equals zero. The score statistic, which is identical to Pearson's goodness-of-fit statistic, is given by

$$X^2 = \gamma \frac{(y_1 - y_2/\gamma)^2}{y_0}.$$

For the sake of completeness we note that the Wald statistic is

$$W^2 = \frac{(y_1 - y_2/\gamma)^2}{(y_1 + y_2/\gamma^2)}.$$

In the following we will not further consider the Wald test. It has been demonstrated by Ng and Tang [3] that the Wald test performs poorly in the present situation, except for $\gamma = 1$, in which case the Wald test is identical to the score test.

Under the null hypothesis both the likelihood-ratio statistic and the score statistic have asymptotically a chi-squared distribution with one degree of freedom. Because the null hypothesis is one-sided the signed versions of the likelihood-ratio test and the score test must be used. That means the tests are applied only if

$$y_2 < \gamma y_1.$$

The critical value for the test statistic is the $(1 - 2\alpha)$ -quantile of the chi-squared distribution.

Conditioning on the total number of observed events y_0 the number of events in either group is binomially distributed. For example

$$y_2 \sim \text{Bin}(\theta, y_0) \quad \text{where } \theta = \frac{\gamma}{1 + \gamma}$$

That means, the p -value of the exact conditional test is

$$p = F(y_2; \theta, y_0)$$

where F denotes the cumulative distribution function of the binomial distribution with success probability θ and sample size y_0 .

With today's high speed computing machines the operating characteristics of the above tests can be easily calculated exactly by summing the probabilities of all single observations in the critical region. The sample space can be visualized as the first quadrant of the plane (cf. Figure 1 below). The evaluation of the test statistic starts at the origin (zero events), proceeds down the y_1 -axis and up the y_2 -axis and stops if the remainder of the sample space has a negligible probability. The procedure will be illustrated in the next section.

A key feature of the test statistics is their monotonicity:

- If (y_1, y_2) is a point of the critical region then both $(y_1 + 1, y_2)$ and $(y_1, y_2 + 1)$ are also points of the critical region.

Sometimes this condition is called convexity of the critical region. In fact, any test lacking this property would contradict common sense. It can be shown analytically that the above tests share the monotonicity property. This helps expediting the power computations considerably.

4 Graphical illustration of the computerized power calculation

For the purpose of illustration we assume a non-inferiority margin ρ of 3.0 and equal total observation times in the two groups so that γ is also equal to 3.0. The lower part of the critical region of the score test at a nominal size of 0.025 is shown in the following figure.

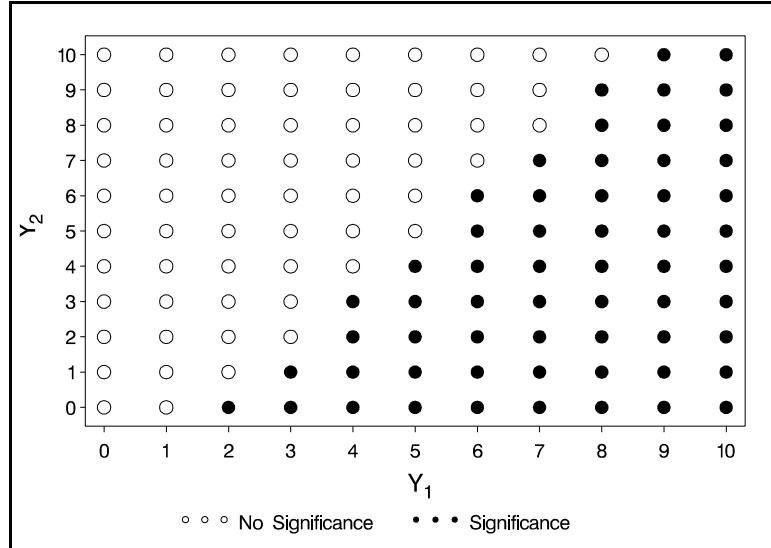


Figure 1: Critical region for $\gamma = \rho = 3.0$, score test, $\alpha = 0.025$

The value of the operating characteristic β at a given parameter vector (μ_1, μ_2) is the sum of the probability of all points (y_1, y_2) that fall into the critical region R :

$$\beta(\mu_1, \mu_2) = \sum_{(y_1, y_2) \in R} \Pr(Y_1 = y_1 | \mu_1) \cdot \Pr(Y_2 = y_2 | \mu_2)$$

In a computerized calculation the summation may start in the column above $y_1 = 1$, i.e. from (1,0) to (1,3). Each point has to be checked for significance and, if significant, its probability has to be added to the operating characteristic. Then the next column is evaluated from (2,0) to (2,6), then column 3 from (3,0) to (3,9) and so on column by column from $(y_1, 0)$ to $(y_1, \gamma y_1)$. Because of the monotonicity it is not necessary to evaluate the test statistic and the probability distribution function at each single point. Suppose, for example, one has found that point (5,4) belongs to the critical region and that (5,5) is outside the critical region. From the monotonicity of the test statistic it follows that all points above (5,5) are also outside the critical region and that one can proceed with column 6. Further, it follows that all points from (6,0) to (6,4) are inside the critical region and need not be checked again for significance.

It is obvious that in this way only a small fraction of all points needs to be considered. These points are displayed in Figure 2. The summation may stop when the probability of the remaining sample space is negligible. For large mean values λ_i the computation time can be notably shortened if the upper and lower tails of the distributions of Y_1 and Y_2 are ignored altogether. If a probability mass of δ is excluded on either side of either distribution then the total error of the calculated operating characteristic can be made less than 2δ provided the computational rounding error is less than $2\delta^2$. Further improvements of the algorithm may be possible, though.

It is a welcome feature of the above algorithm that it can be easily modified to allow for size and power calculations of different tests simultaneously. To be specific, Figure 3 below displays the critical regions of all three tests, the likelihood-ratio test, the score test, and the exact conditional test for the illustrative example with $\gamma = \rho = 3.0$.

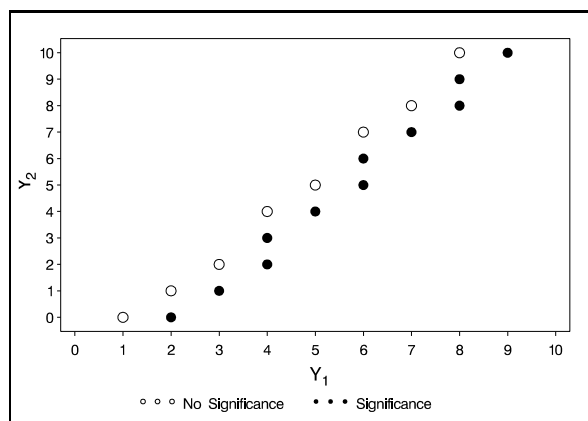


Figure 2: Points to be checked

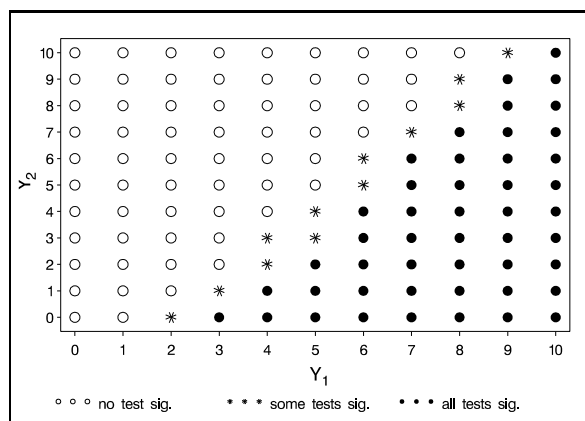


Figure 3: Critical region of all three tests

In the computerized calculation one needs to keep track of the maximum y_2 value such that (y_1, y_2) is in the critical region of all tests. For $y_1 = 6$ in Figure 3 this maximum is $y_2 = 4$. When starting with the next column at $y_1 = 7$ one can add the cumulative distribution function from $(7,0)$ to $(7,4)$ to the operating characteristic of all three tests. The points above $(7,4)$ are then checked separately for the three tests until all tests are non-significant, i.e. until point $(7,8)$.

A realization of this algorithm in a SAS data step is provided at the end of this paper. To give an idea of the computation time we note that with SAS 8.02 under Windows 2000 the computations for Figures 4 and 5 together took 0.3 seconds. For $\mu_1 = \mu_2 = 2$, $\rho = 1.01$, and $m_1 = m_2 = 10^5$ the computations took 0.7 seconds.

5 An example

Suppose a clinical trial is planned to show that a new treatment is not inferior to a standard treatment in the prevention of infections. Non-inferiority would be accepted if the infection rate under the new treatment is not more than 1.5 times the infection rate under the standard treatment. It is assumed that the number of infections follows a Poisson distribution. Each patient should be followed up for one year. For power calculations the average infection rate under the standard treatment is estimated to be 2.0 per year. The one-sided significance level is set at 0.025. The following two graphics show the size and the power of the three tests for sample sizes between 30 and 100 per group. Figure 4 illustrates that the likelihood-ratio test meets the nominal size of 0.025 very good. The score test is only slightly liberal. The exact conditional test is conservative as was to be expected for this type of test, similar to Fisher's exact test for two by two tables. Figure 5 shows that the power of the exact conditional test is not much lower than the power of the other two tests, particularly for power values above 0.9. Under the assumptions made above a sample size of 65 per group would provide a power

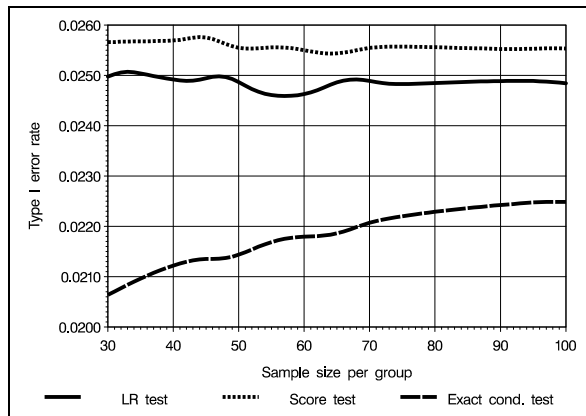


Figure 4: Comparison of type I error rate

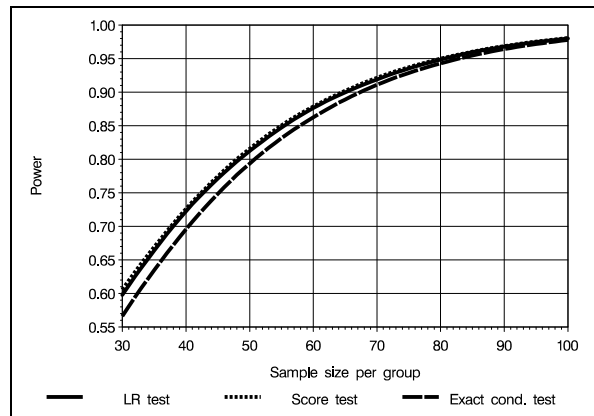


Figure 5: Comparison of power

of 0.90 for both, the likelihood-ratio test and the score test, and a power of 0.89 for the exact conditional test.

The power cannot be improved by using unequal sample sizes. This is illustrated in Figure 6 below.

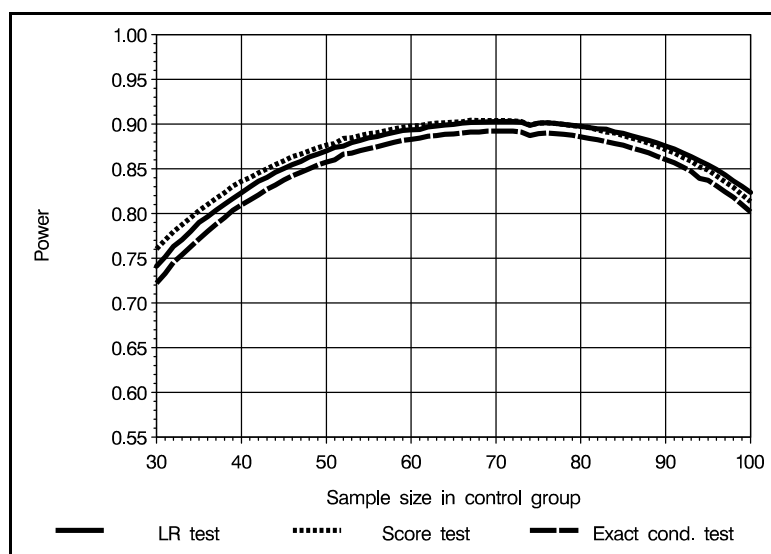


Figure 6: Power of the tests depending on the splitting of the total sample size of 130

The power of the likelihood-ratio test ranges between 0.896 and 0.903 for all values of n_1 between 62 and 81. The power of the exact conditional test is between 0.884 and 0.892.

However, the sample size may be chosen to minimize the maximum possible type I error rate. The following two graphics show how the test size depends on the true mean value μ_1 for two different sample size combinations.

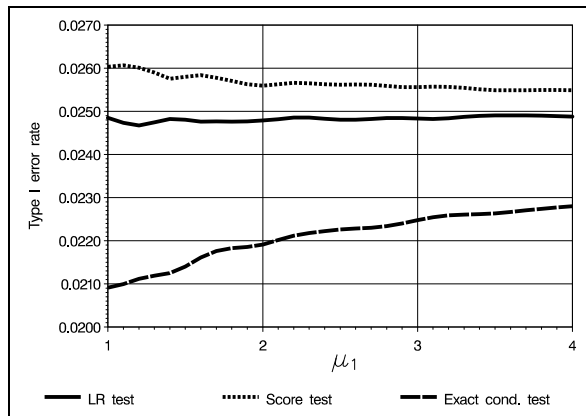


Figure 7: Size of tests if $n_1 = 63$, $n_2 = 67$

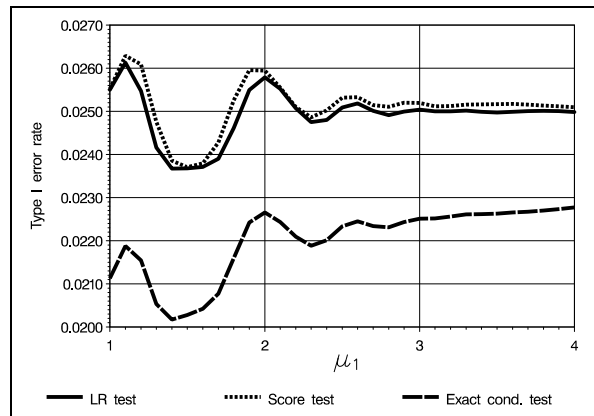


Figure 8: Size of tests if $n_1 = 74$, $n_2 = 56$

Obviously the type I error rate of the likelihood-ratio test is perfectly maintained for mean values μ_1 between 1 and 4 if the sample size in the first group is 63 (Figure 7). A sample size of 74 in the first group leads to a minor inflation of the type I error rate for mean values μ_1 around 1.0 and 2.0 (Figure 8).

We close with some general experiences that may be verified in particular applications using the attached program. In non-inferiority trials, in which ρ is greater than one, the likelihood ratio test controls the nominal size typically better than the score test. The sample size ratio should be chosen such that γ lies between 1 and 1.5.

For superiority studies with ρ equal to 1.0 the score test controls the type I error rate slightly better than the likelihood-ratio test. Equal sample sizes are a good choice in this case. With equal sample sizes the actual size of the score test is less than 0.0255 provided the total sample size times the mean value μ_1 is at least 33. Usually one may not want to use smaller sample sizes because the power would be too low. In fact, in most applications the sample size required for a power of 0.9 will be high enough to use the exact conditional test without relevant loss in power and without any compromise regarding the test size. This is practically relevant considering Figures 7 and 8 because in clinical trials the actual sample size is often somewhat different from the planned sample size.

6 Conclusions

An exact method for calculating sample size and power of three statistical tests comparing two Poisson distributions was introduced. For the realization in SAS only SAS Base is required. The monotonicity of the tests facilitates the calculation in short time. This makes the use of approximate formulae or simulations redundant.

It was shown in an example how the exact calculation of size and power can lead to an optimal determination of the sample sizes for the two groups. These calculations should not be driven too far, however. After all, the accuracy of the calculated operating characteristic depends crucially on the adequacy of the Poisson distribution and this will always remain an unprovable assumption.

References

- [1] The European Agency for the Evaluation of Medical Products (1998); ICH Topic E9, *Statistical principles for clinical trials*, <http://www.emea.eu.int/pdfs/human/ich/036396en.pdf>
- [2] Roehmel, J., Hauschke, D., Koch, A., Pigeot, I. (2005); Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren. Nicht-Unterlegenheit in klinischen klinischen Studien. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 48, 562-571
- [3] Ng, H.K.T., Tang, M.-L. (2005); Testing the equality of two Poisson means using the rate ratio. *Statistics in Medicine*, 24, 955-965

Contact information

Corinna Miede
Accovion GmbH
Softwarecenter 3
35037 Marburg
Germany
Phone: ++49 +6421 9 48 49 28
E-mail: Corinna.Miede@accovion.com
www.accovion.com

Jochen Mueller-Cohrs
Accovion GmbH
Softwarecenter 3
35037 Marburg
Germany
Phone: ++49 +6421 9 48 49 27
E-mail: Jochen.Mueller-Cohrs@accovion.com
www.accovion.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

SAS program

```
TITLE "Power for comparing two Poisson distributions: H0: MU2/MU1 > RHO";

* Input parameters (to be supplied by the user);
* ERR: tolerated error for the power calculation;
* (ignoring machine dependent rounding errors);
* ALPHA, RHO, MU1, MU2, M1, M2 (see Text);

* Output parameters;
* POW_LRT: power likelihood-ratio test;
* POW_SCO: power score test;
* POW_ECT: power exact conditional test;

* Further parameters and variables;
* GAM, LAM1, LAM2 denote GAMMA, LAMBDA1 and LAMBDA2 (see text);

* (Y1,Y2) are values of the sample space;
* Y2X is maximum y2 such that (y1,y2) is signif. for all tests at current y1;
* CDF1 is the cumulative distribution function of Y1;
* CDF2 is the cumulative distribution function of Y2X;
* NOSIG is 1 if no test is significant for the current (y1,y2), and 0 otherwise;

data a(keep=alpha rho--m2 pow_lrt--pow_ect);

err=1E-6;                * user input;
del=err/2;
eps=1-del;
ini=-del+2*del**2;

do alpha = 0.025;        * user input;
    crit = cinv(1-2*alpha,1);
do rho = 1.5;            * user input;
do mu1 = 2;              * user input;
do mu2 = mu1*rho, mu1;   * user input;
do m1 = 30 to 100;       * user input;
do m2 = m1;              * user input;

    gam=(m2/m1)*rho;
    theta=gam/(1+gam);
    lam1=m1*mu1;
    lam2=m2*mu2;

    pow_lrt=ini;
    pow_sco=ini;
    pow_ect=ini;
```

```

y1=0;
cdf1=pdf("Poisson",0,lam1);
p1=pdf("Poisson",1,lam1);
do while(cdf1+p1<del);
    y1+1;
    cdf1+p1;
    p1=pdf("Poisson",y1+1,lam1);
end;

y2x=-1;
cdf2=0;
p2=pdf("Poisson",0,lam2);
do while(cdf2+p2<del);
    y2x+1;
    cdf2+p2;
    p2=pdf("Poisson",y2x+1,lam2);
end;

do until(cdf1>eps | cdf2>eps);

    y1+1;
    p1=pdf("Poisson",y1,lam1);
    cdf1+p1;
    py=p1*cdf2;

    pow_lrt+py;
    pow_sco+py;
    pow_ect+py;

    y2=y2x;
    do until(nosig);
        y2+1;
        p2=pdf("Poisson",y2,lam2);
        py=p1*p2;
        y0=y1+y2;

        * Likelihood ratio test;

        ly1=y1*log(y1)-y0*log(y0/(1+gam));
        ly2=0;
        if y2>0 then ly2=y2*log(y2/gam);
        chi_lrt=2*(ly1+ly2);
        sig_lrt=(chi_lrt>crit);
        pow_lrt+(py*sig_lrt);
    end;
end;

```

```

* Score test;

chi_sco=gam*(y1-y2/gam)**2/y0;
sig_sco=(chi_sco>crit);
pow_sco+(py*sig_sco);

* Exact conditional test;

pval_ect=cdf("binom",y2,theta,y0);
sig_ect=(pval_ect<alpha);
pow_ect+(py*sig_ect);

if sig_lrt & sig_sco & sig_ect then do;
    y2x=y2;
    cdf2+p2;
end;

nosig=(y2>y1*gam)+(1-sig_lrt)*(1-sig_sco)*(1-sig_ect);
end;

end;

py=1-cdf1;
pow_lrt+py;
pow_sco+py;
pow_ect+py;
output;

end; end; end; end; end; end;
run;

proc print data=a;
    by alpha--mu2 notsorted;
    id alpha--mu2 ;
    pageby mu2;
    format pow_lrt--pow_ect 6.4;
run;

```