Optimal exact tests for multiple binary endpoints

Robin Ristl

Section for Medical Statistics, Center of Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna

joint work with
Dong Xi, Ekkehard Glimm and Martin Posch

Padua, April 2016

Reasons to analyse multiple endpoints

- Some diseases need more than one endpoint for full characterization
- Discriminating between endpoints with and without treatment effect may help to understand mode of action of a drug
- If we want to show just some effect, testing multiple endpoints can increase power compared to single endpoint test (Senn and Bretz, 2007)

What is the problem with small sample sizes?

- Asymptotic distibution may not reflect true distributions with sufficient accuracy
- Low precision of nuisance parameter estimates
- Limited model complexity/risk of overfitting
- Low power

Exact tests provide some solution

- A test is exact, if the null distribution of its test statistic is perfectly known
- Often achieved by conditioning on sufficient statistics or by permutation
- Alleviates problem of inaccurate asymptotics and nuisance paramters
- Drawback: Exact tests typically have discrete test statistics and therefore can become overly conservative

We will focus on one-sided tests for multiple binary endpoints.

Fisher’s Exact Test

- Test statistic T is the number of successes in the treatment group
- Conditional on margins, T has hypergeometric distr.

Example: n=15 per group, marginal number of successes = 20, one-sided test

Null distribution of T

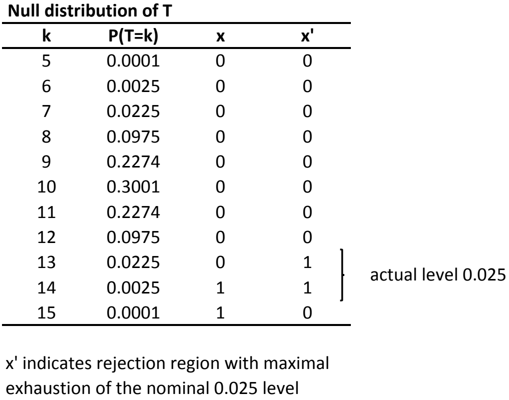
k	**P(T=k)**	**x**
5	0.0001	0
6	0.0025	0
7	0.0225	0
8	0.0975	0
9	0.2274	0
10	0.3001	0
11	0.2274	0
12	0.0975	0
13	0.0225	0
14	0.0025	1
15	0.0001	1

actual level 0.0026

x indicates usual rejection region

A simple idea of optimizing exhaustion of nominal level

Select a set R of points in the support of T, such that $P_{H_0}(R)$ is as close as possible to α, but not greater than α.



Null distribution of T

k	**P(T=k)**	**x**	**x'**
5	0.0001	0	0
6	0.0025	0	0
7	0.0225	0	0
8	0.0975	0	0
9	0.2274	0	0
10	0.3001	0	0
11	0.2274	0	0
12	0.0975	0	0
13	0.0225	0	1
14	0.0025	1	1
15	0.0001	1	0

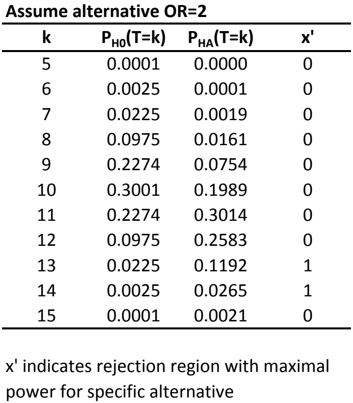
actual level 0.025

x' indicates rejection region with maximal exhaustion of the nominal 0.025 level

Perfect alpha exhaustion, but the most extreme point is not in R. Motivation for additional "no holes" condition!

Optimizing for a specific alternative (Paroush 1969)

Select R, such that under the alternative H_A, $P_{H_A}(R)$ is maximal, subject to $P_{H_0}(R) \leq \alpha$.



Assume alternative OR=2

k	$P_{H0}(T=k)$	$P_{HA}(T=k)$	x'
5	0.0001	0.0000	0
6	0.0025	0.0001	0
7	0.0225	0.0019	0
8	0.0975	0.0161	0
9	0.2274	0.0754	0
10	0.3001	0.1989	0
11	0.2274	0.3014	0
12	0.0975	0.2583	0
13	0.0225	0.1192	1
14	0.0025	0.0265	1
15	0.0001	0.0021	0

x' indicates rejection region with maximal power for specific alternative

Again, there is a "hole" in the rejection region, which makes it inpracticable.

Two binary endpoints

Marginal null hypotheses H_1 and H_2, vector of test statistics $T = (T_1, T_2)$. Want to test $H_1 \cap H_2$ (no effect in any EP), H_1 and H_2. How should we select the (marginal) rejection regions?

Null distribution of T_1

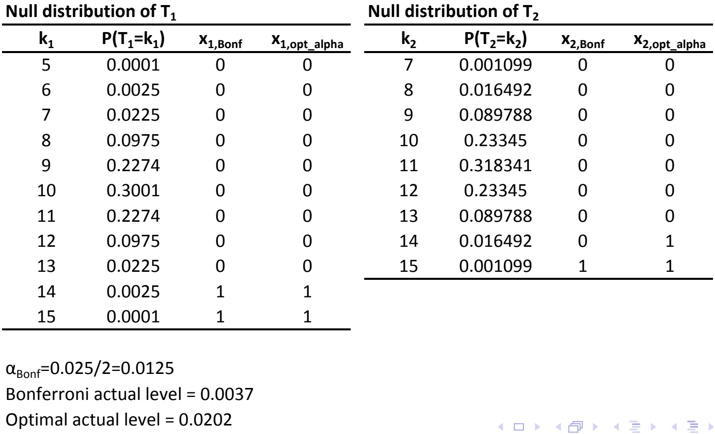
k_1	$P(T_1=k_1)$
5	0.0001
6	0.0025
7	0.0225
8	0.0975
9	0.2274
10	0.3001
11	0.2274
12	0.0975
13	0.0225
14	0.0025
15	0.0001

Null distribution of T_2

k_2	$P(T_2=k_2)$
7	0.001099
8	0.016492
9	0.089788
10	0.23345
11	0.318341
12	0.23345
13	0.089788
14	0.016492
15	0.001099

Optimal Bonferroni tests using marginal distributions

Select marginal regions R_1 and R_2, such that $P_{H_0}(R_1) + P_{H_0}(R_2) \to \max$, subject to $P_{H_0}(R_1) + P_{H_0}(R_2) \leq \alpha$ and "no holes". Leads to optimally weighted Bonferroni.



Null distribution of T_1

k_1	$P(T_1=k_1)$	$x_{1,Bonf}$	x_{1,opt_alpha}
5	0.0001	0	0
6	0.0025	0	0
7	0.0225	0	0
8	0.0975	0	0
9	0.2274	0	0
10	0.3001	0	0
11	0.2274	0	0
12	0.0975	0	0
13	0.0225	0	0
14	0.0025	1	1
15	0.0001	1	1

Null distribution of T_2

k_2	$P(T_2=k_2)$	$x_{2,Bonf}$	x_{2,opt_alpha}
7	0.001099	0	0
8	0.016492	0	0
9	0.089788	0	0
10	0.23345	0	0
11	0.318341	0	0
12	0.23345	0	0
13	0.089788	0	0
14	0.016492	0	1
15	0.001099	1	1

α_{Bonf}=0.025/2=0.0125

Bonferroni actual level = 0.0037

Optimal actual level = 0.0202

General rejection regions R for a global hypothesis on m endpoints using the joint distribution of T

- Idea: Use the joint distribution of $T = (T_1, ..., T_m)$ to define optimal rejection regions
  - Gutman and Hochberg 2007 used linear integer programming, but did not consider “no holes” restriction
- Find the joint distribution as permutation distribution between the groups
  - Similar to minP test (Westfall 1989), but restricted shape of R there
- Note that permutation null hypothesis is exchangeability, which is a stronger hypothesis than $\cap_{i=1}^{m} H_i$.
  - Westfall and Troendle 2008, Klingenberg et al. 2009, give arguments when permutation testing is acceptable
  - Simply put, it is justified if we can be assume that the new treatment is not doing worse than the old treatment in terms of the joint distribution of endpoints.

Optimal tests using the joint distribution of T

Assume that the joint distribution of $T = (T_1, \ldots, T_m)$ under $H_0 = \cap_{i=1}^{m} H_i$ is known. H_0 is rejected if $t \in R$.

To establish type I error control, the rejection region R has to satisfy the condition

(i) $P_{H_0}(T \in R) \leq \alpha$

The "no holes" condition to avoid implausible regions is formalized as

(ii) If $(t_1, \ldots, t_m) \in R$ then

$$\{(s_1, \ldots, s_m) : s_1 \geq t_1, \ldots, s_m \geq t_m\} \subseteq R$$

Different optimization goals

1. Optimize exhaustion of nominal level

   $$R : P_{H_0}(T \in R) \to \max, \text{ s.t. conditions } (i) \text{ and } (ii)$$

2. Optimize the number of elements (points) $|R|$ in the rejection region

   $$R : |R| \to \max, \text{ s.t. conditions } (i) \text{ and } (ii)$$

3. Optimize the power for a specific alternative

   $$R : P_{H_A}(T \in R) \to \max, \text{ s.t. conditions } (i) \text{ and } (ii)$$

Numeric optimization

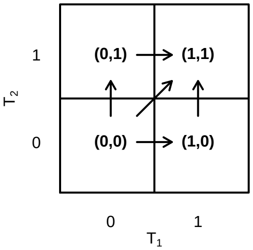
- Can write the problem as linear integer program
- Tried LP-solver lpSolve in R
- Works, but can become numerically unstable or take very long
- Propose a branch and bound algorithm instead
- Our algorithm makes specific use of "no holes" constraint in calculating the upper and lower bounds of a node
- Has reasonable runtime, provides feasible current solution even in case iterations are limited

Branch and bound algorithm

Example search space

t	x
(0,0)	?
(0,1)	?
(1,0)	?
(1,1)	?

Implications due to cond. (ii)



- Branch
  - $x_1 = (1, ?, ?, ?)$
  - $x_0 = (0, ?, ?, ?)$
- Fill due to cond. (ii)
  - $x_1 = (1, 1, 1, 1)$
  - $x_0 = (0, ?, ?, ?)$
- Bound
  - x_1 fully branched
  - Lower bound for x_0: objective value for $x = (0, 0, 0, 0)$
  - Upper bound for x_0: objective value for $x = (0, 1, 1, 1)$
- Remove unfeasible solutions (cond. (i)) and solutions with upper bound $<$ current largest lower bound
- Iterate until there are only fully branched solutions. These are optimal.

Extended example

Example data and assumed alternative

	Observed marginal	Assumed alternative	
	successes	Group 1	Group 2
No success	4	0.1	0.7
EP1 only	4	0.2	0.1
EP2 only	6	0.2	0.1
Both EPs	16	0.5	0.1

n per group = 15

Solutions in the example - Optimal alpha exhaustion

Optimal alpha exhaustion

T_2	0	0.2	2.2	9.7	22.7	30	22.7	9.7	2.2	0.2	0	
15					0	0	0	0	0	0	0	0.1
14				0	0.2	0.5	0.5	0.3	0.1	0	0	1.6
13			0	0.4	1.5	2.7	2.6	1.3	0.4	0	0	9
12		0	0.3	1.7	4.7	7	6	2.9	0.7	0.1	0	23.3
11	0	0.1	0.7	3.1	7.2	9.6	7.2	3.1	0.7	0.1	0	31.8
10	0	0.1	0.7	2.9	6	7	4.7	1.7	0.3	0		23.3
9	0	0	0.4	1.3	2.6	2.7	1.5	0.4	0			9
8	0	0	0.1	0.3	0.5	0.5	0.2	0				1.6
7	0	0	0	0	0	0	0					0.1
	5	6	7	8	9	10	11	12	13	14	15	

P(R) = 2.488 %, |R| = 18

T_1

Solutions in the example - Optimal area

Maximal number of elements in R

T_2

	0	0.2	2.2	9.7	22.7	30	22.7	9.7	2.2	0.2	0	
15					0	0	0	0	0	0	0	0.1
14				0	0.2	0.5	0.5	0.3	0.1	0	0	1.6
13			0	0.4	1.5	2.7	2.6	1.3	0.4	0	0	9
12		0	0.3	1.7	4.7	7	6	2.9	0.7	0.1	0	23.3
11	0	0.1	0.7	3.1	7.2	9.6	7.2	3.1	0.7	0.1	0	31.8
10	0	0.1	0.7	2.9	6	7	4.7	1.7	0.3	0		23.3
9	0	0	0.4	1.3	2.6	2.7	1.5	0.4	0			9
8	0	0	0.1	0.3	0.5	0.5	0.2	0				1.6
7	0	0	0	0	0	0	0					0.1
	5	6	7	8	9	10	11	12	13	14	15	

P(R) = 2.377 %, |R| = 23

T_1

Solutions in the example - Optimal power

Optimal power

T_2	0	0	0	0	0.1	1.2	7.7	24.6	37	24.2	5.2	
15					0	0.3	1.7	4.9	5.6	2.4	0.3	15.2
14				0	0	0.5	3.2	10.3	15.2	9.1	1.6	39.8
13			0	0	0	0.3	2.1	7	12.1	9.1	2.2	32.9
12		0	0	0	0	0.1	0.6	2.1	3.7	3.2	1	10.7
11	0	0	0	0	0	0	0.1	0.3	0.5	0.4	0.1	1.4
10	0	0	0	0	0	0	0	0	0	0		0.1
9	0	0	0	0	0	0	0	0	0			0
8	0	0	0	0	0	0	0	0				0
7	0	0	0	0	0	0	0					0
	5	6	7	8	9	10	11	12	13	14	15	

P(R) = 86.45 %, |R| = 20

T_1

Closed testing procedure and consonance

- Closed test: A null hypothesis is rejected, if all intersection hypothesis it is contained in are rejected by a local level α test.
    - Provides family wise type I error control at level α
- Can use locally optimal tests for all intersection hypotheses in the closed test
    - For two binary EPs: After rejecting $H_1 \cap H_2$, use marginal Fisher tests for H_1 and H_2 at level α.
- Consonance property: If $\cap_{i=1}^{m} H_i$ is rejected, also at least one marginal H_i can be rejected.
    - For 2 EPs, our procedure can be easily constrained to be consonant
    - For more endpoints, joint optimization of the full closed test required, computationally intense

Example - Optimal alpha exhaustion with consonance

Optimal alpha exhaustion

T_2

	0	**0.2**	**2.2**	**9.7**	**22.7**	**30**	**22.7**	**9.7**	**2.2**	**0.2**	**0**	
15					**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.1**
14				**0**	**0.2**	**0.5**	**0.5**	**0.3**	**0.1**	**0**	**0**	**1.6**
13			**0**	**0.4**	**1.5**	**2.7**	**2.6**	**1.3**	**0.4**	**0**	**0**	**9**
12		**0**	**0.3**	**1.7**	**4.7**	**7**	**6**	**2.9**	**0.7**	**0.1**	**0**	**23.3**
11	**0**	**0.1**	**0.7**	**3.1**	**7.2**	**9.6**	**7.2**	**3.1**	**0.7**	**0.1**	**0**	**31.8**
10	**0**	**0.1**	**0.7**	**2.9**	**6**	**7**	**4.7**	**1.7**	**0.3**	**0**		**23.3**
9	**0**	**0**	**0.4**	**1.3**	**2.6**	**2.7**	**1.5**	**0.4**	**0**			**9**
8	**0**	**0**	**0.1**	**0.3**	**0.5**	**0.5**	**0.2**	**0**				**1.6**
7	**0**	**0**	**0**	**0**	**0**	**0**	**0**					**0.1**
	5	6	7	8	9	10	11	12	13	14	15	

P(R) = 2.009 %, |R| = 22

T_1

Alpha consistency

- Alpha consistency: If the test rejects at level α, it also rejects at all levels $\alpha' > \alpha$.
- It is not required for a valid test at a pre-specified level, but it is necessary to define p-values
- The optimal tests are not alpha consistent
- The rejection region for an alpha consistent test can be found by a simple greedy algorithm:
- Out of all points that are feasible in terms of the "no holes" condition, always adds the point with the smallest contribution to the type I error rate

Example - Alpha consistent test

Alpha consistent test

T_2

	0	**0.2**	**2.2**	**9.7**	**22.7**	**30**	**22.7**	**9.7**	**2.2**	**0.2**	**0**	
15					**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.1**
14				**0**	**0.2**	**0.5**	**0.5**	**0.3**	**0.1**	**0**	**0**	**1.6**
13			**0**	**0.4**	**1.5**	**2.7**	**2.6**	**1.3**	**0.4**	**0**	**0**	**9**
12		**0**	**0.3**	**1.7**	**4.7**	**7**	**6**	**2.9**	**0.7**	**0.1**	**0**	**23.3**
11	**0**	**0.1**	**0.7**	**3.1**	**7.2**	**9.6**	**7.2**	**3.1**	**0.7**	**0.1**	**0**	**31.8**
10	**0**	**0.1**	**0.7**	**2.9**	**6**	**7**	**4.7**	**1.7**	**0.3**	**0**		**23.3**
9	**0**	**0**	**0.4**	**1.3**	**2.6**	**2.7**	**1.5**	**0.4**	**0**			**9**
8	**0**	**0**	**0.1**	**0.3**	**0.5**	**0.5**	**0.2**	**0**				**1.6**
7	**0**	**0**	**0**	**0**	**0**	**0**	**0**					**0.1**
	5	6	7	8	9	10	11	12	13	14	15	

P(R) = 2.377 %, |R| = 23

T_1

Example for unconditional power - Setting

- Sample size per group is $n = 15$, nominal one-sided significance level is 2.5%.
- Success probability for both endpoints is 0.735 in the treatment group and 0.265 in the control group.
- For a single marginal Fisher test, this means a power of 60%.
- The correlation between the endpoints is 0.5

Example for unconditional power - Some results

Power to reject the indicated hyptheses following a closed test.

Global test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2
Bonferroni	64.0	64.0	42.5	53.3	53.3
Bonferroni opt. alpha	74.7	74.7	44.6	60.0	59.2
minP	75.7	75.7	44.5	60.1	60.0
Optimal alpha	83.6	74.3	44.6	59.5	59.4
Optimal area	83.4	76.1	44.6	60.4	60.2
Optimal power	84.6	75.4	44.6	60.0	59.9
Consonant opt. power	76.2	76.2	44.6	60.4	60.4
Alpha consistent	82.9	76.2	44.6	60.4	60.4

Observations from numeric power study

- Optimized tests are far more powerful than Bonferroni, and also better than minP
- Enforcing consonance seems not important here (less power for global test, only small power gain for elementary rejection)
- Alpha consistent test shows stable performance close to the optimal test
- Optimal power test can be far better than other optimal tests, but often these get close
- Multiplicity for free? The power to reject a specific elementary H_i is hardly reduced compared to that of a single Fisher test.

Conclusions

- Optimizing the multivariate rejection region offers a notable advantage over more simple methods
- In the small sample setting, where this approach can have the greatest impact, optimal solutions are found within short computation times
- Application may require additional planning effort, careful prespecification in study protocol
- Worthwhile, if the aim is to make best use of multiple exact hypotheses tests from a small data sample

Literature

S. Senn and F. Bretz. Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics, 6(3):161-170, 2007. doi: 10.1002/pst.301.

Ronald A Fisher. The logic of inductive inference. Journal of the Royal Statistical Society, pages 39-82, 1935.

Jacob Paroush. Integer programming technique to construct statistical tests. The American Statistician, 23(5):43-44, 1969.

R. Gutman and Y. Hochberg. Improved multiple test procedures for discrete dis- tributions: New ideas and analytical review. Journal of Statistical Planning and Inference, 137(7):2380-2393, 2007. doi: 10.1016/j.jspi.2006.08.006.

Peter H Westfall and S Stanley Young. P value adjustments for multiple tests in multivariate binomial models. Journal of the American Statistical Association, 84 (407):780-786, 1989.

Peter H Westfall and James F Troendle. Multiple testing with minimal assumptions. Biometrical Journal, 50(5):745-755, 2008.

B Klingenberg, A Solari, L Salmaso, and F Pesarin. Testing marginal homogeneity against stochastic order in multivariate ordinal data. Biometrics, 65(2):452-462, 2009.