Better power calculations for preliminary two sample proportion comparisons

(Optimising animal use in vaccine research and development)

Damian Collins and Paul Hick (NSW DPI)

Elizabeth Macarthur Agricultural Institute, Camden NSW

September 2nd 2024

Motivation: Preliminary vaccine studies

- Emerging vaccine technologies (mRNA and modified viruses): tremendous potential for improved animal health
  - Early "proof-of-concept" studies: screen many candidate vaccines – determine the best to pursue through R&D pathways.
- Want to know that a candidate vaccine induces some protective immune response
  - 70% efficacy is considered economically viable
  - so we want to compare a disease rate of 97.5% (controls) say vs 30% (vaccinated)
    - we want sufficient power to be able to detect a difference like that
  - so... looking for a **big** effect of the vaccine with a **small** sample size
- Animal ethics (AEC) require optimized studies for a meaningful study outcome
  - i.e. sufficient power & minimum no of animals
  - further pressure to minimise sample size because they are emergency animal diseases that require high levels of biocontainment

Motivation: Preliminary vaccine studies

- Emerging vaccine technologies (mRNA and modified viruses): tremendous potential for improved animal health
  - Early "proof-of-concept" studies: screen many candidate vaccines – determine the best to pursue through R&D pathways.
- Want to know that a candidate vaccine induces some protective immune response
  - 70% efficacy is considered economically viable
  - so we want to compare a disease rate of 97.5% (controls) say vs 30% (vaccinated)
    - we want sufficient power to be able to detect a difference like that
  - so... looking for a **big** effect of the vaccine with a **small** sample size
- Animal ethics (AEC) require optimized studies for a meaningful study outcome
  - i.e. sufficient power & minimum no of animals
  - further pressure to minimise sample size because they are emergency animal diseases that require high levels of biocontainment

Motivation: Preliminary vaccine studies

- Emerging vaccine technologies (mRNA and modified viruses): tremendous potential for improved animal health
  - Early "proof-of-concept" studies: screen many candidate vaccines – determine the best to pursue through R&D pathways.
- Want to know that a candidate vaccine induces some protective immune response
  - 70% efficacy is considered economically viable
  - so we want to compare a disease rate of 97.5% (controls) say vs 30% (vaccinated)
    - we want sufficient power to be able to detect a difference like that
  - so... looking for a **big** effect of the vaccine with a **small** sample size
- Animal ethics (AEC) require optimized studies for a meaningful study outcome
  - i.e. sufficient power & minimum no of animals
  - further pressure to minimise sample size because they are emergency animal diseases that require high levels of biocontainment

NSW GOVERNMENT

“Off the shelf” power calculations for 2 proportions

- either software (R/GenStat), online calculator or even a book (!)

"Off the shelf" power calculations for 2 proportions

- either software (R/GenStat), online calculator or even a book (!)
- three main approximations

No Yates correction		*Yates correction*
Approx v1	*Approx v2 (better)*	*Approx v3*
- Snedecor Cochran (1989), p 129	- power.prop.test (R)	- Sokal and Rolfe (1981) p 766
- Power and Sample Size website	- SBNtest (Genstat)	- propTestPower (EnvStats in R)
- Select Statistics website		- Epitools website

Example: 97.5% vs 30%, n=6 and 9

- Power given n:
  - say n=6 or n=9 per group - big differences especially for n=6

	No correction		*Yates corr.*
	Approx 1	*Approx 2*	*Approx 3*
n=6	93%	75%	43%
n=9	99%	92%	77%

Example: 97.5% vs 30%, n=6 and 9

- Power given n:
  - say n=6 or n=9 per group - big differences especially for n=6

	No correction		*Yates corr.*
	Approx 1	*Approx 2*	*Approx 3*
n=6	93%	75%	43%
n=9	99%	92%	77%

- n for given power (rounded up)

	No correction				*Yates corr.*	
	Approx 1		*Approx 2*		*Approx 3*	
Power=80%	4	(4.0)	7	(6.6)	10	(9.4)
Power=90%	6	(5.4)	9	(8.4)	12	(11.1)
Power=99%	10	(9.5)	14	(13.3)	17	(16.1)

  - quite a big difference!

Some more “exact” alternatives

- Fisher’s exact test is the first thought for a more “exact” alternative

Some more “exact” alternatives

- Fisher’s exact test is the first thought for a more “exact” alternative
- some implementations now
  - power.fisher.test (statmod in R) (simulation approach)
  - Robin Ristl (University of Vienna (2024)) webpage
    - exact power and sample size calculations using the Fisher test
    - written in Javascript (!!)
  - G*Power (thanks Steve Morris!)

Some more "exact" alternatives

- Fisher's exact test is the first thought for a more "exact" alternative
- some implementations now
  - power.fisher.test (statmod in R) (simulation approach)
  - Robin Ristl (University of Vienna (2024)) webpage
    - exact power and sample size calculations using the Fisher test
    - written in Javascript (!!)
  - G*Power (thanks Steve Morris!)
- however Fisher's exact test is not the only "exact" way to compare two proportions
  - there is the Barnard test and similar (e.g. Boschloo)
  - the whole controversy about conditional (e.g. Fisher) vs unconditional (e.g. Barnard) tests...
    - Fay. M and Hunsberger, SA (2021)
    - Ripamonti, E, Lloyd, C. and Quatto, P. (2017)
  - conditional tests (e.g. Fisher) are more conservative than unconditional (e.g. Barnard)

Example: 97.5% vs 30%, n=6 and 9 (with Fisher)

- power for n=6 or n=9 per group - Fisher most like corrected

	No correction		*Yates corr.*	**Fisher**
	Approx 1	*Approx 2*	*Approx 3*	
n=6	93%	75%	43%	***38%***
n=9	99%	92%	77%	***86%***

- n for given power - Fisher closest to corrected

	No correction				*Yates corr.*		**Fisher**
	Approx 1		*Approx 2*		*Approx 3*		***rounded***
Power=80%	4	(4.0)	7	(6.6)	10	(9.4)	***9***
Power=90%	6	(5.4)	9	(8.4)	12	(11.1)	***11***
Power=99%	10	(9.5)	14	(13.3)	17	(16.1)	***16***

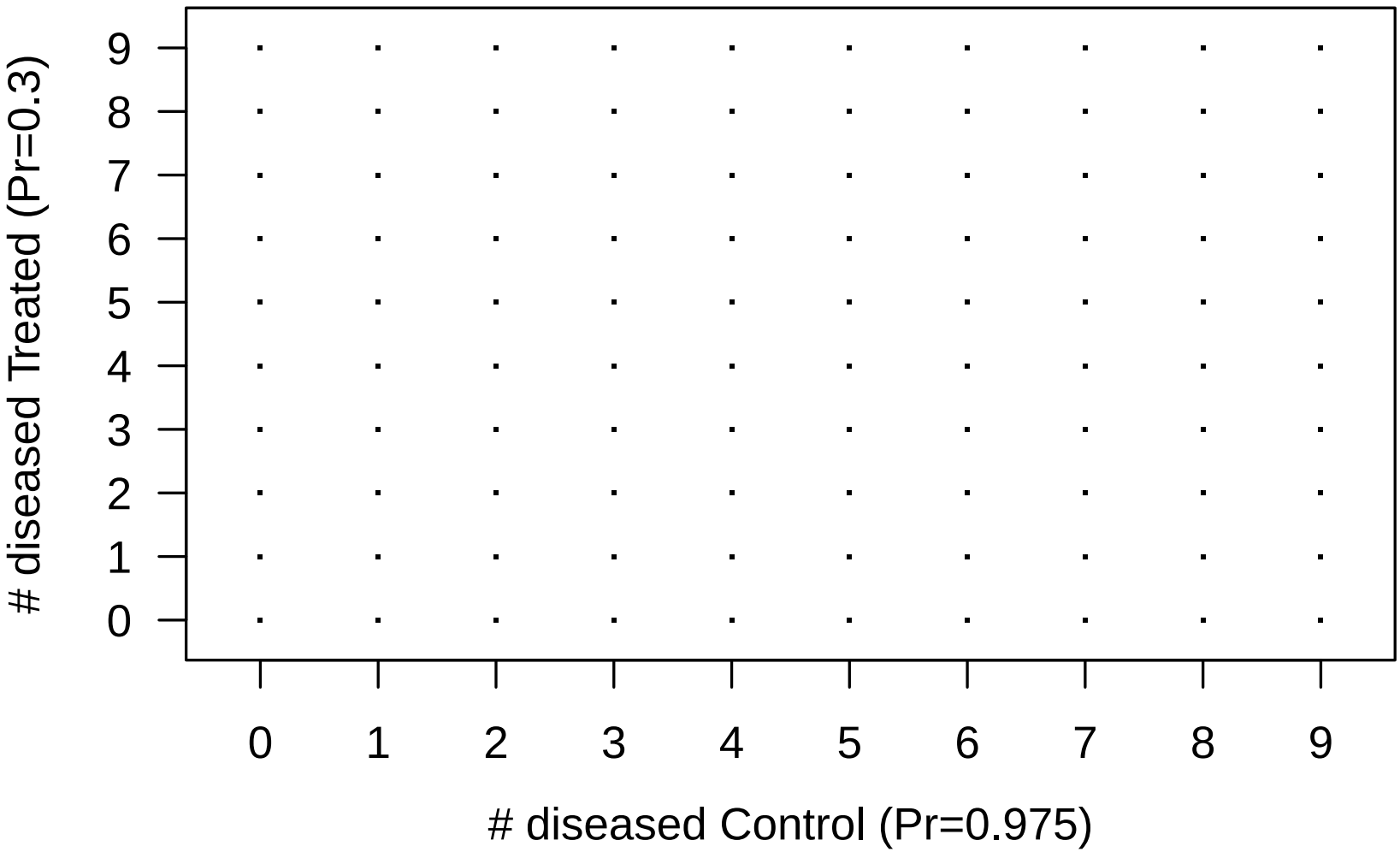
  - quite a big difference!

My idea

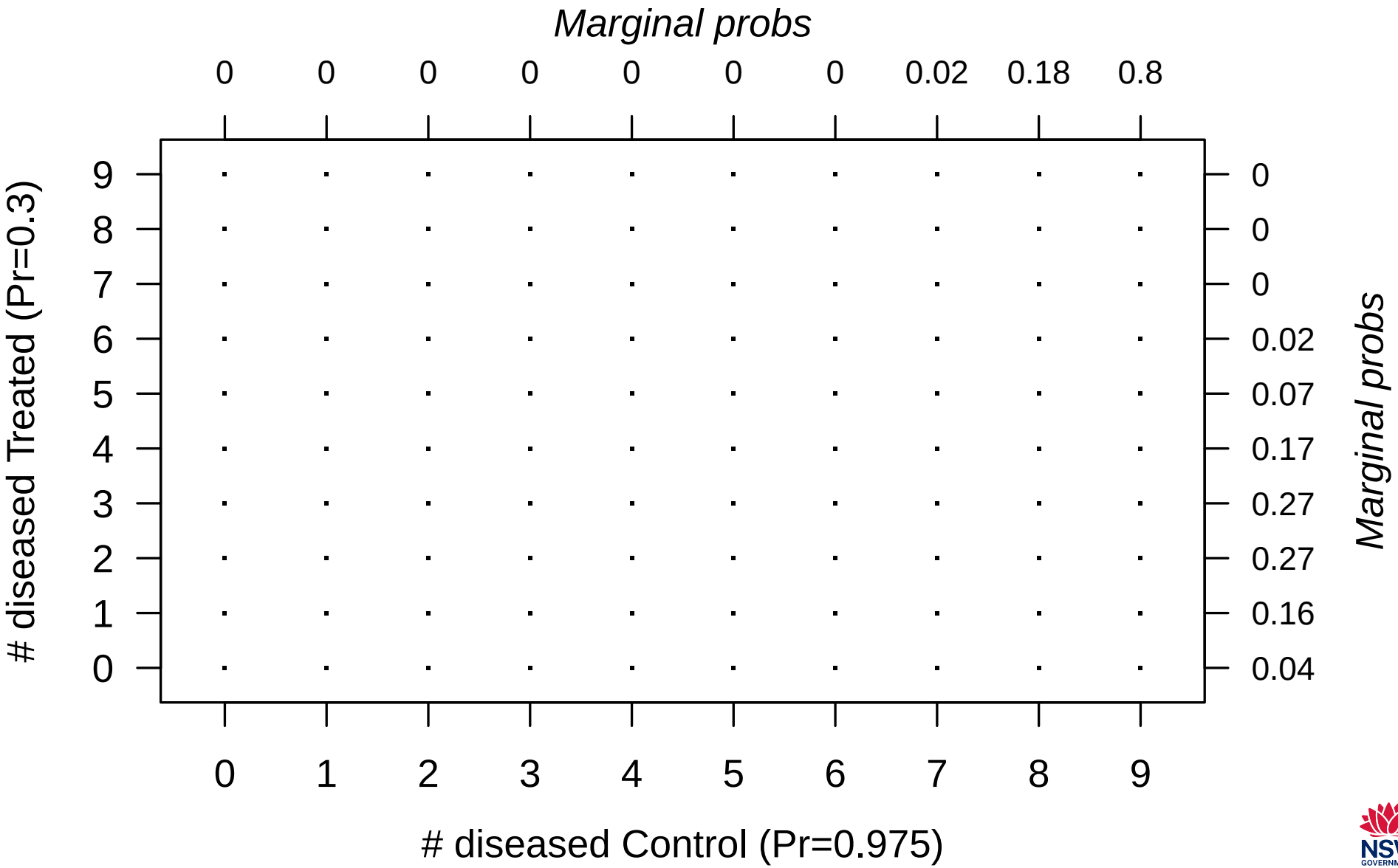
- since it is a small sample space...
  - simply enumerate all the possible combinations
  - and determine which are significant

- show what I mean for $n_1 = n_2 = 9$
  - and $p_C = 0.975$ (control) vs $p_T = 0.3$ (treatment)

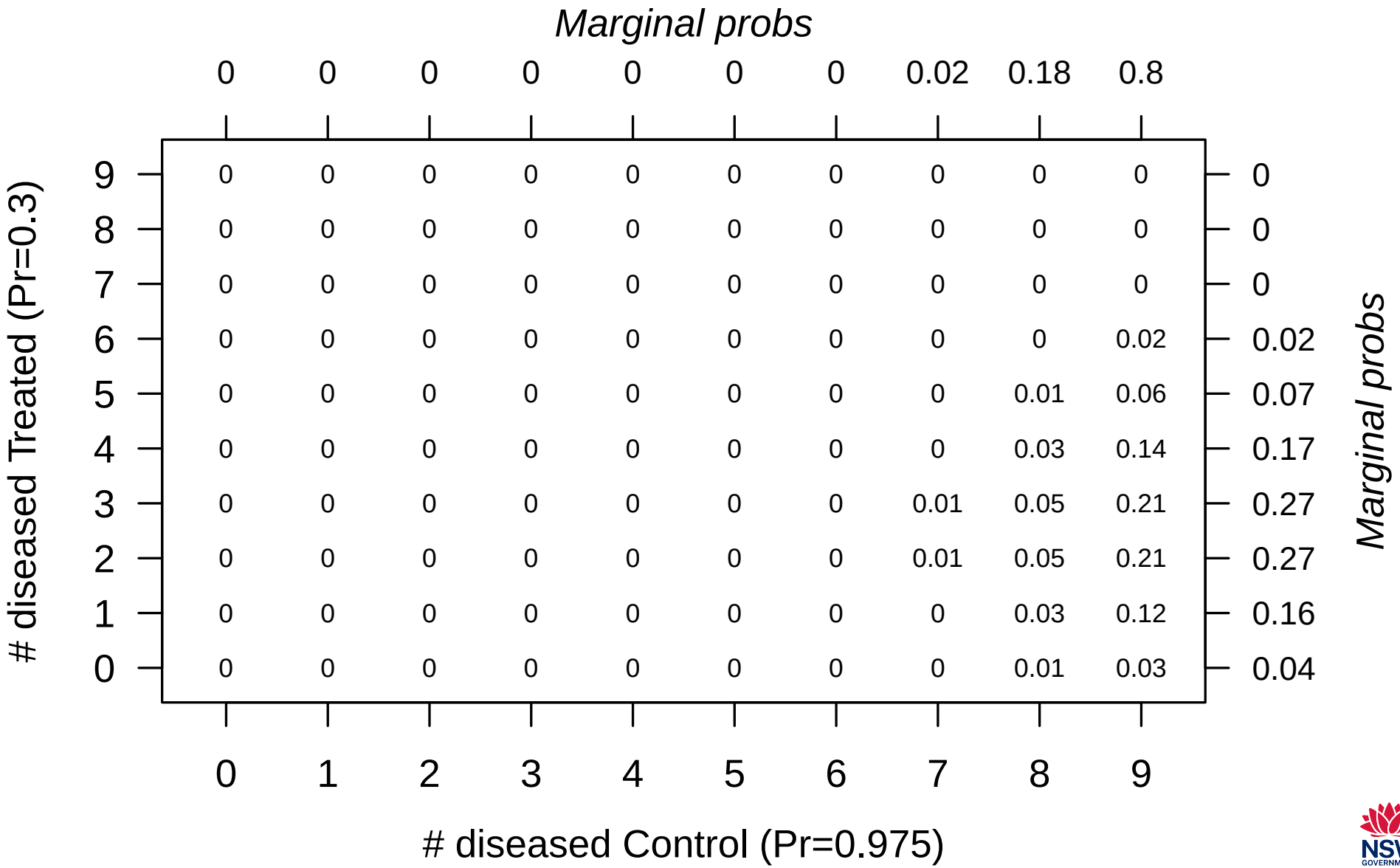
A diagram - all the 10^2=100 possibilities



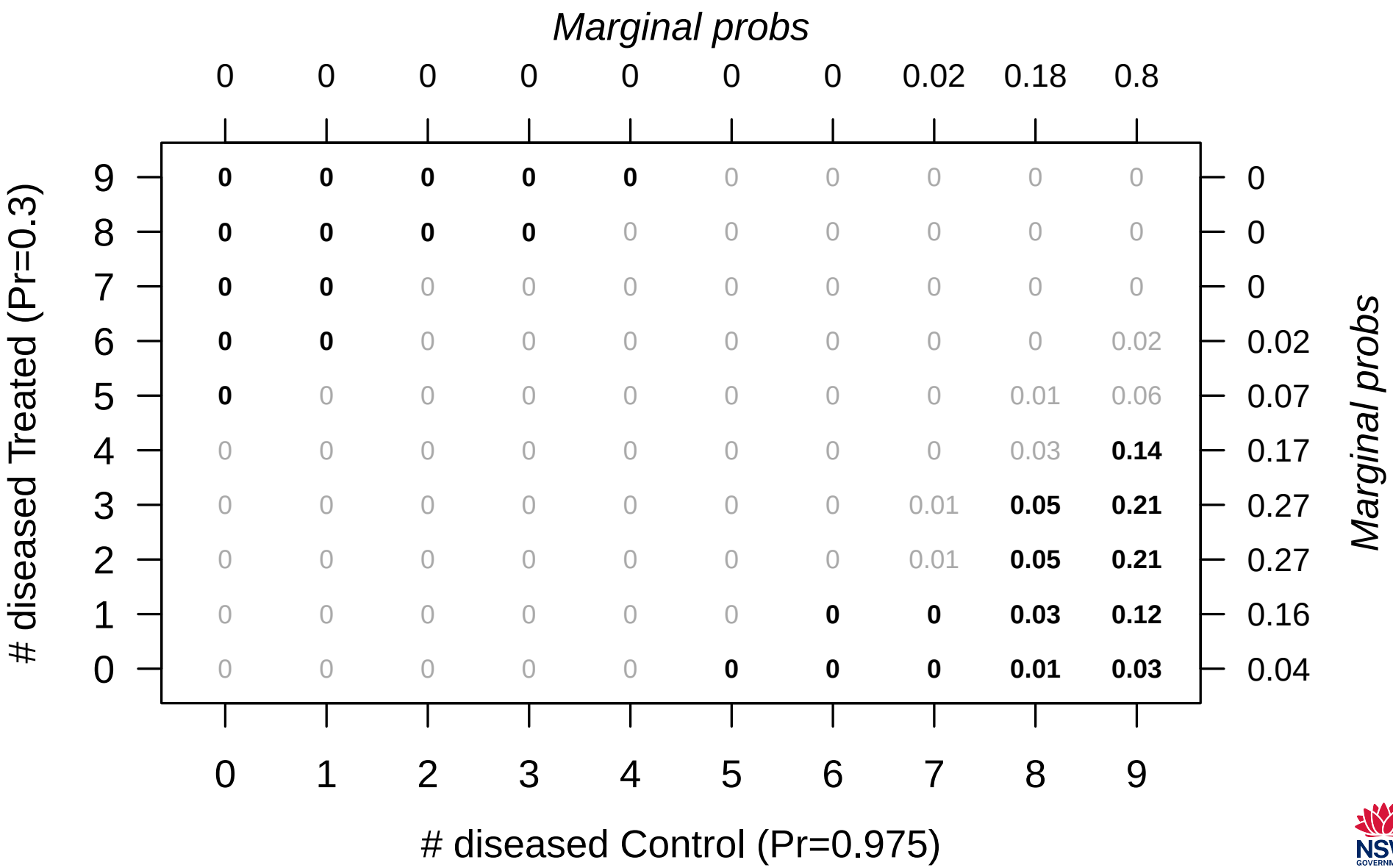
.... the binomial probabilities



... the joint probabilities



... significant (using Fisher test) in bold



Some R code

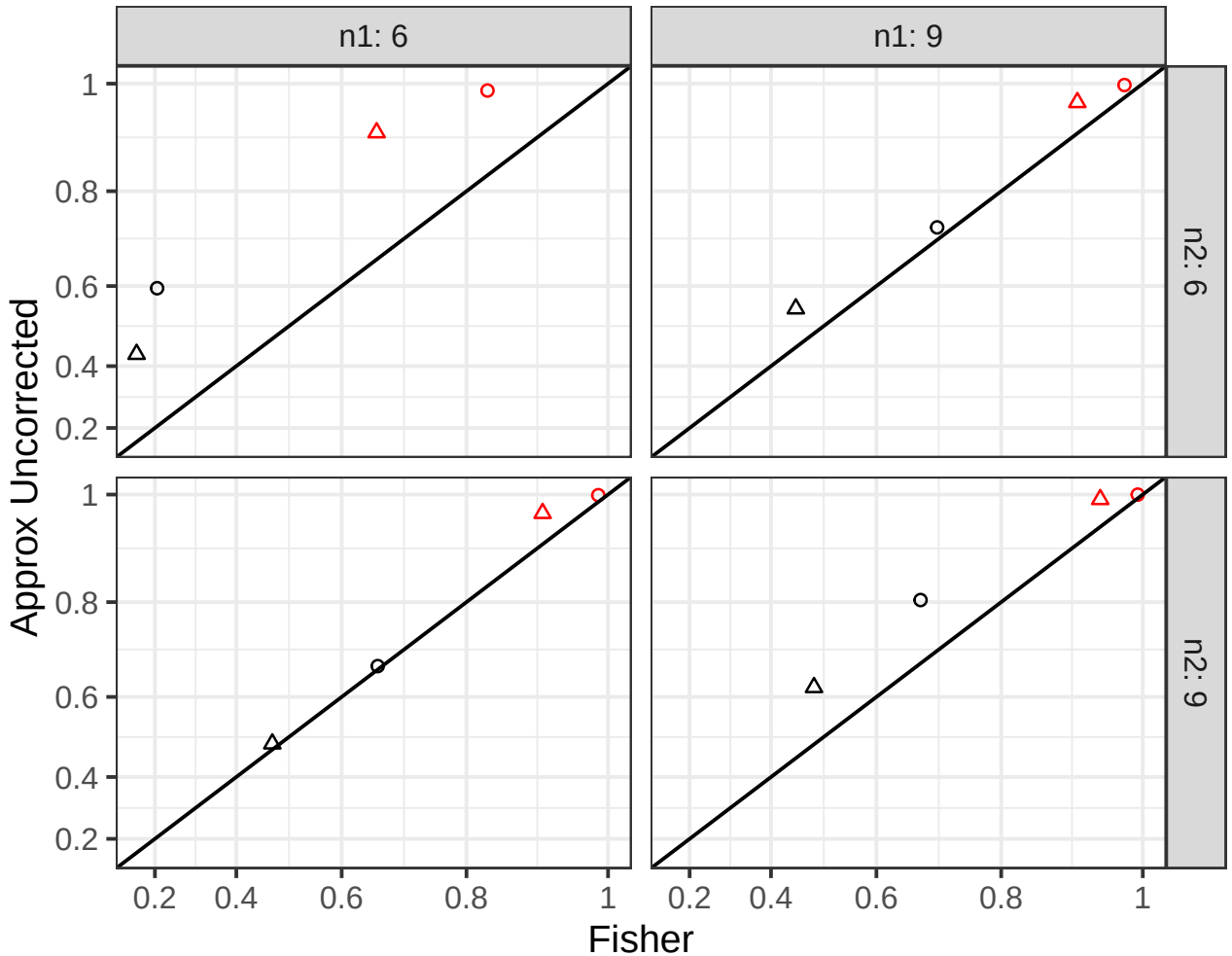
```
## set up dataframe of all 100 outcomes 0...9 x 0...9
df <- expand.grid(T1=0:9, T2=0:9)
##
## calculate the probability of each outcome -
## product of two binomial probabilities
df$prob <- dbinom(df$T1,9,0.025) * dbinom(df$T2,9,0.7)
##
## define the function to obtain the p.value
testfn <- function(x,y) fisher.test(cbind(c(x,y),
       9-c(x,y)))$p.val
##
## determine the p-value for each outcome (using mapply)
df$pval <- mapply(testfn, df$T1, df$T2)
##
## report the power at the console nearest %
cat("power=",round(100*sum(df$prob[df$pval<0.05])),"%")
```

Some comparisons

- comparing my method with Fisher & Barnard vs approximate power calculations
  - approx power calculations
    - v2 (better) (`power.prop.test`)
    - v3 (corrected) (`propTestPower` in `EnvStats` package)
  - my method with the Fisher or Barnard test
    - Barnard test using Barnard package
- for the combinations of the following parameters
  - $n_1 = 6, 9$
  - $n_2 = 6, 9$
  - $p_1 = 0.975, 0.9$ (control)
  - $p_2 = 0.4, 0.1$ (treatment)
- so over n_1, n_2, p_1 and p_2 there are $2^4 = 16$ combinations

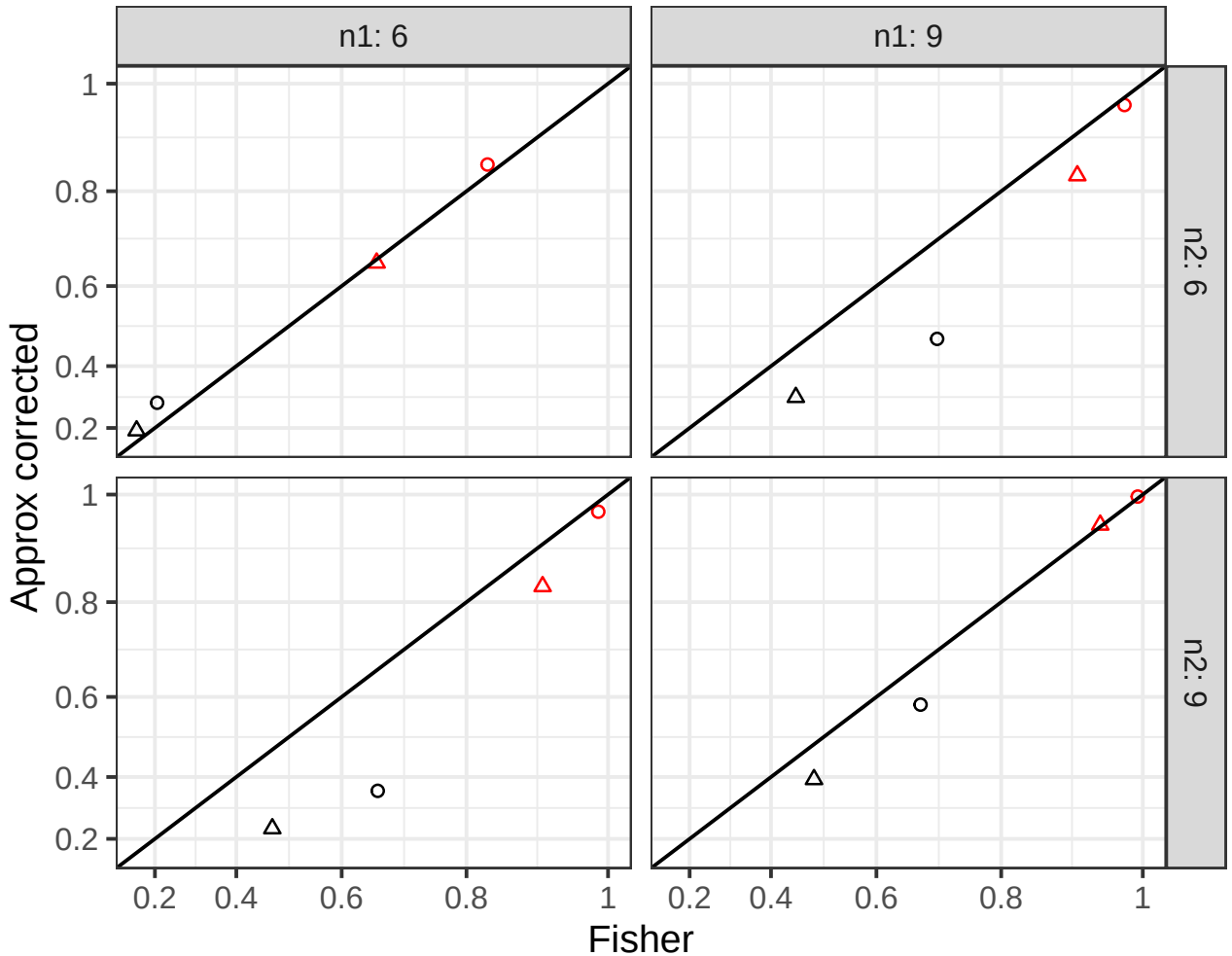
Fisher vs approx (uncorrected)

As expected!



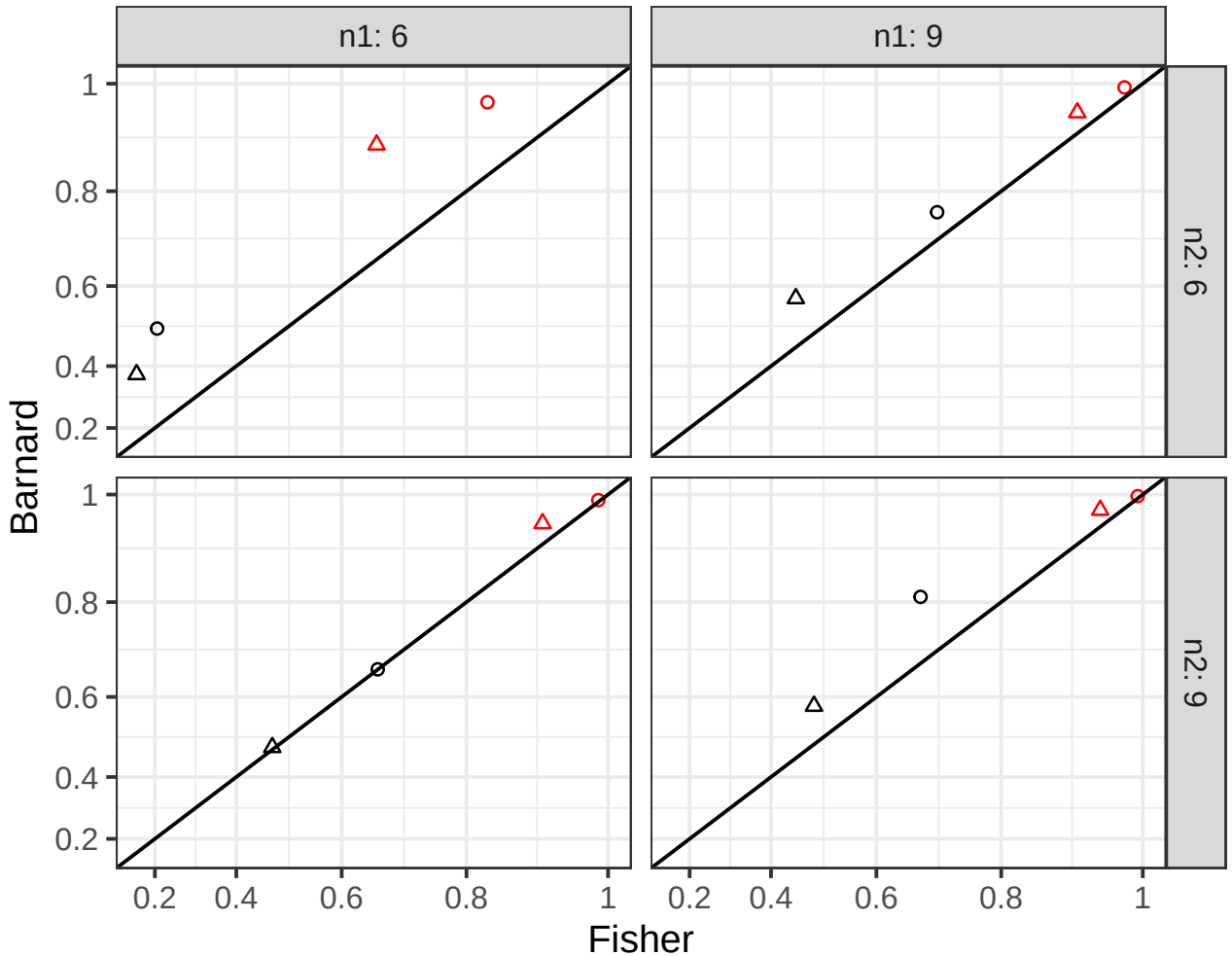
Fisher vs approximate corrected test

Yates correction overcorrects... (as expected)



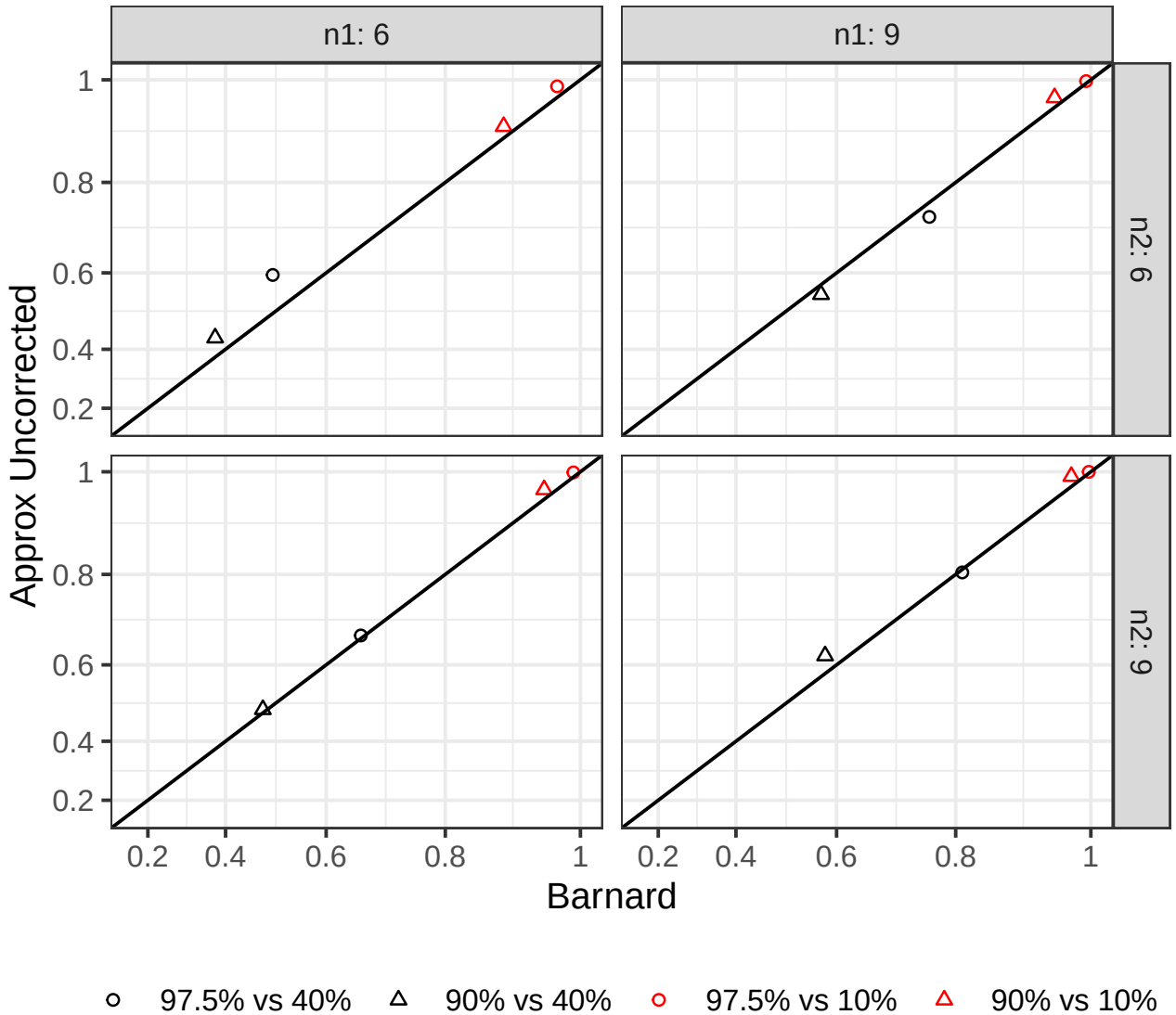
Fisher vs Barnard

Barnard more powerful (as expected) espec. with lower n (more discrete).



Barnard vs approx uncorrected

They almost match (unexpected!)



Issues

- my approach: simply a better way of calculating power for small sample proportion studies
  - why not? more choice and flexibility...
  - even for larger sample sizes???

Issues

- my approach: simply a better way of calculating power for small sample proportion studies
  - why not? more choice and flexibility...
  - even for larger sample sizes???
- my approach – which test to use?
  - conditional (e.g. Fisher) or unconditional (e.g. Barnard)???
  - Fisher is more conservative (and well-known), so a pragmatic approach (precautionary principle) might be to use that.

Issues

- my approach: simply a better way of calculating power for small sample proportion studies
  - why not? more choice and flexibility...
  - even for larger sample sizes???
- my approach – which test to use?
  - conditional (e.g. Fisher) or unconditional (e.g. Barnard)???
  - Fisher is more conservative (and well-known), so a pragmatic approach (precautionary principle) might be to use that.
- sample size for given power?
  - my method calculates power for given sample sizes
  - need brute force method (like the U Vienna webpage)

Side remarks about power calculations

- general confusion about power options (and on the website calculators)

Side remarks about power calculations

- general confusion about power options (and on the website calculators)
- what power do you really need? Is 80% enough?

Side remarks about power calculations

- general confusion about power options (and on the website calculators)
- what power do you really need? Is 80% enough?
- what is the difference do we really need/want to detect?
  - often unknown or unclear - reverse-engineered to match a fixed sample size...

Side remarks about power calculations

- general confusion about power options (and on the website calculators)
- what power do you really need? Is 80% enough?
- what is the difference do we really need/want to detect?
  - often unknown or unclear - reverse-engineered to match a fixed sample size...
- and should it be a one-sided test? Is that cheating????
  - people usually do two-sided tests (because that's often the default)