Statistical Power Calculations

Manuel AR Ferreira

Massachusetts General Hospital
Harvard Medical School
Boston

Boulder, 2007

Outline

1. Aim
2. Statistical power
3. Estimate the power of linkage / association analysis
   - Analytically
   - Empirically
4. Improve the power of linkage analysis

1. Aim

1. Know what type-I error and power are

2. Know that you can/should estimate the power of your linkage/association analyses (analytically or empirically)

3. Know that there a number of tools that you can use to estimate power

4. Be aware that there are MANY factors that increase type-I error and decrease power

2. Statistical power

H_0: Person A is not guilty

H_1: Person A is guilty – send him to jail

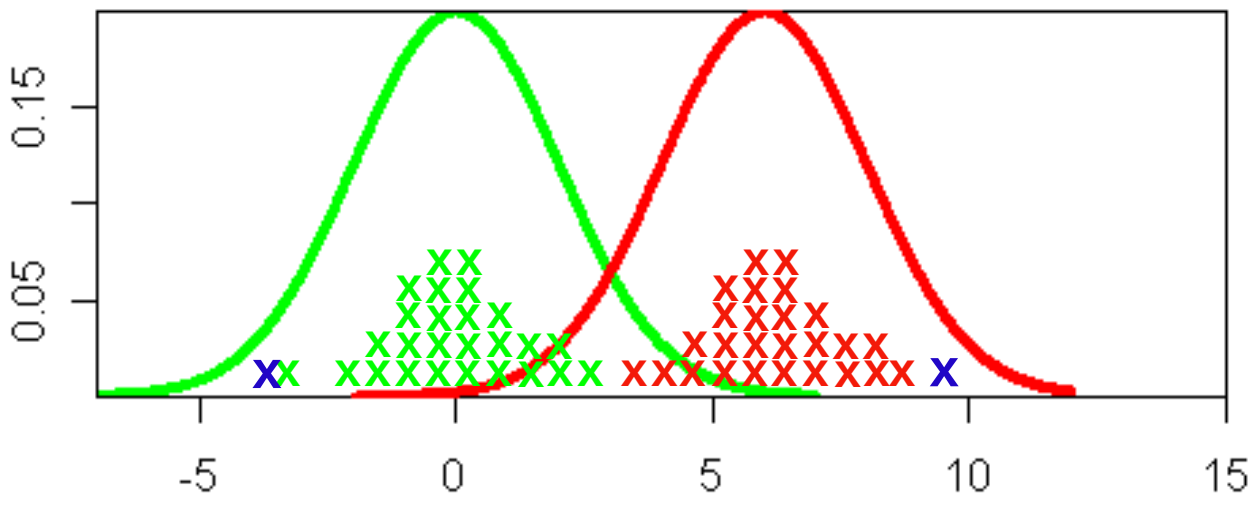
		In reality…	
		H_0 is true	**H_1 is true**
We decide…	**H_0 is true**	$1 - \alpha$	β Type-2 error
	H_1 is true	α Type-1 error	$1 - \beta$ Power

Power: probability of declaring that something is true when in reality it is true.

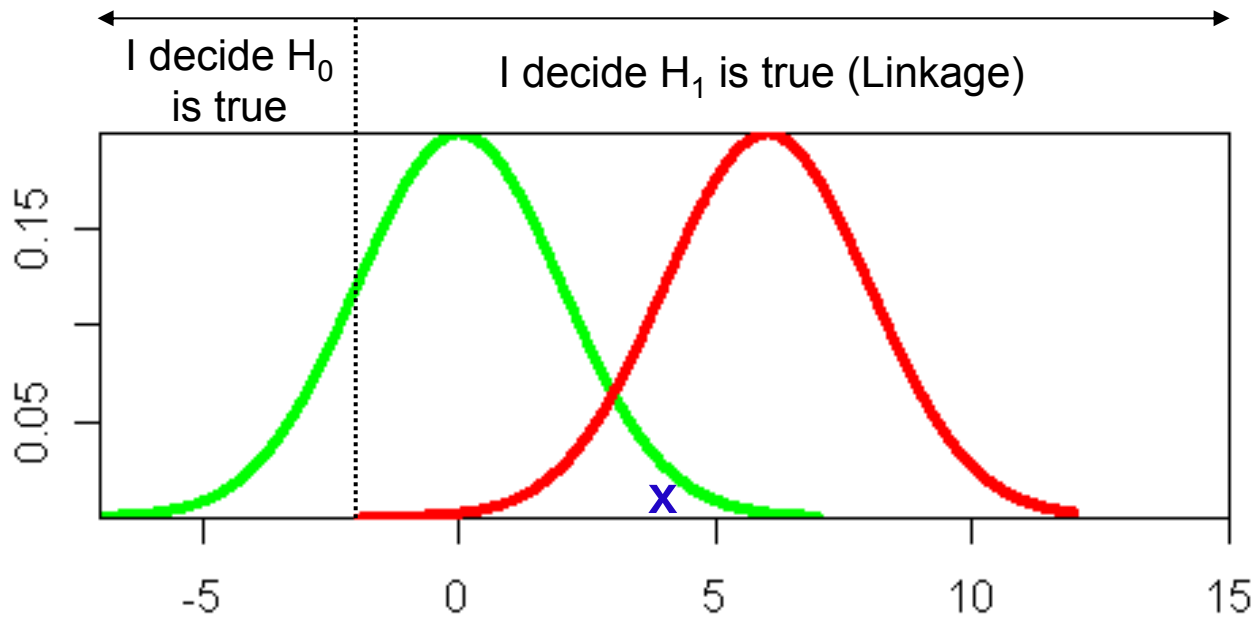
H_0: There is <u>NO</u> linkage between a marker and a trait

H_1: There is linkage between a marker and a trait

Linkage test statistic has different distributions under H_0 and H_1

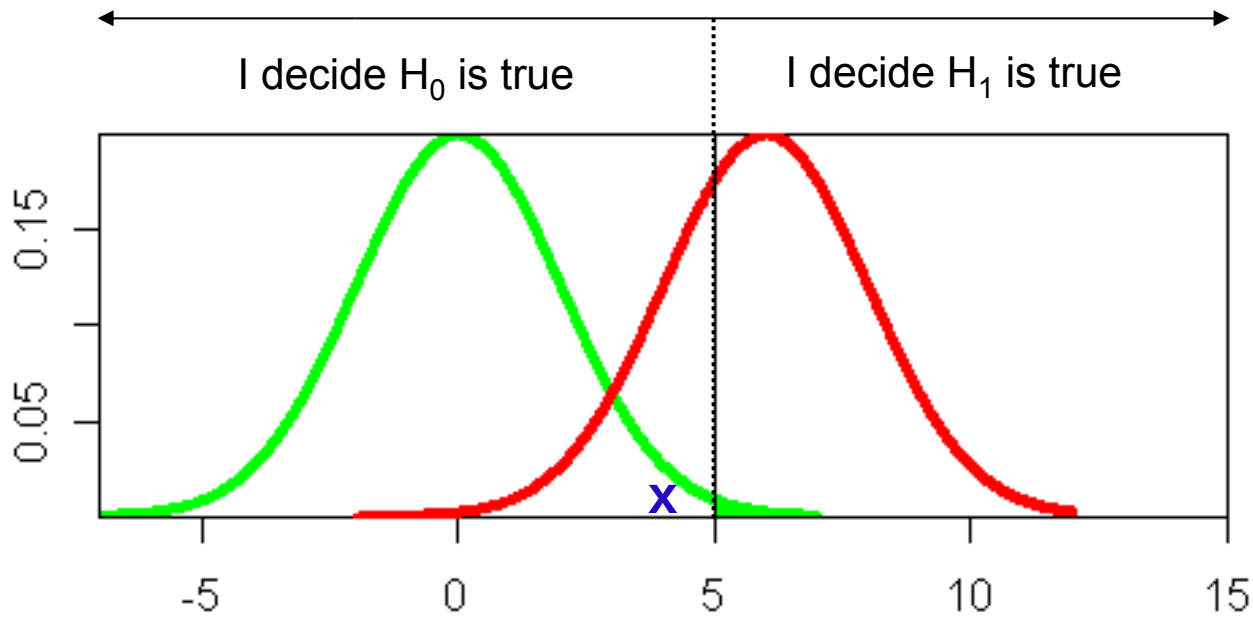


Where should I set the threshold to determine significance?



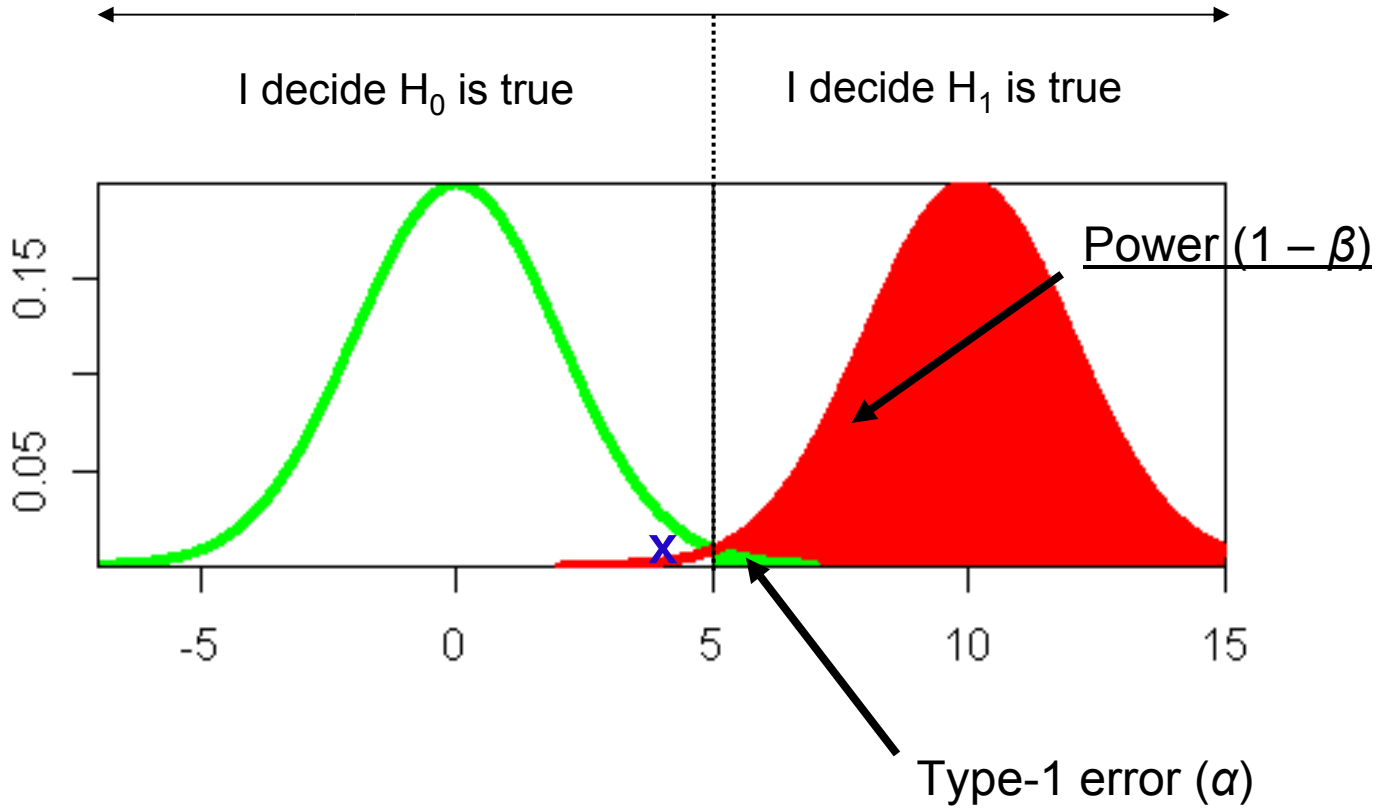
Threshold	Power ($1 - \beta$)	Type-1 error (α)
To low	**High**	**High**

Where should I set the threshold to determine significance?



Threshold	Power ($1 - \beta$)	Type-1 error (α)
To low	**High**	**High**
To high	**Low**	**Low**

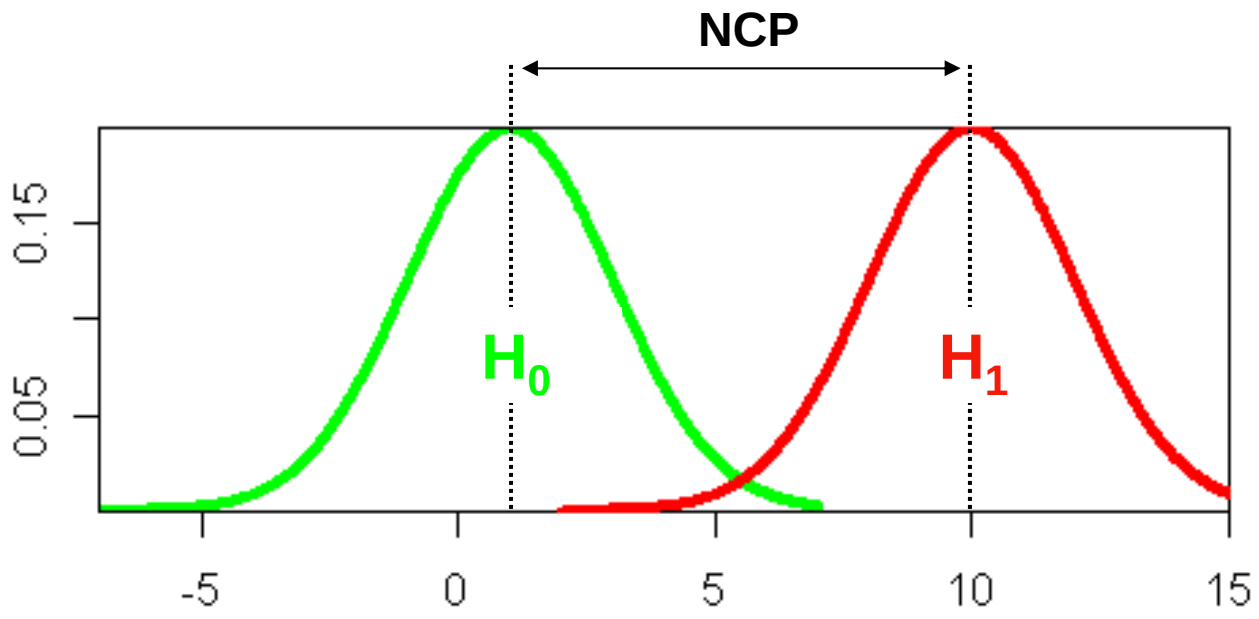
How do I maximise Power while minimising Type-1 error rate?



1. Set a high threshold for significance (i.e. results in low α [e.g. 0.05-0.00002])

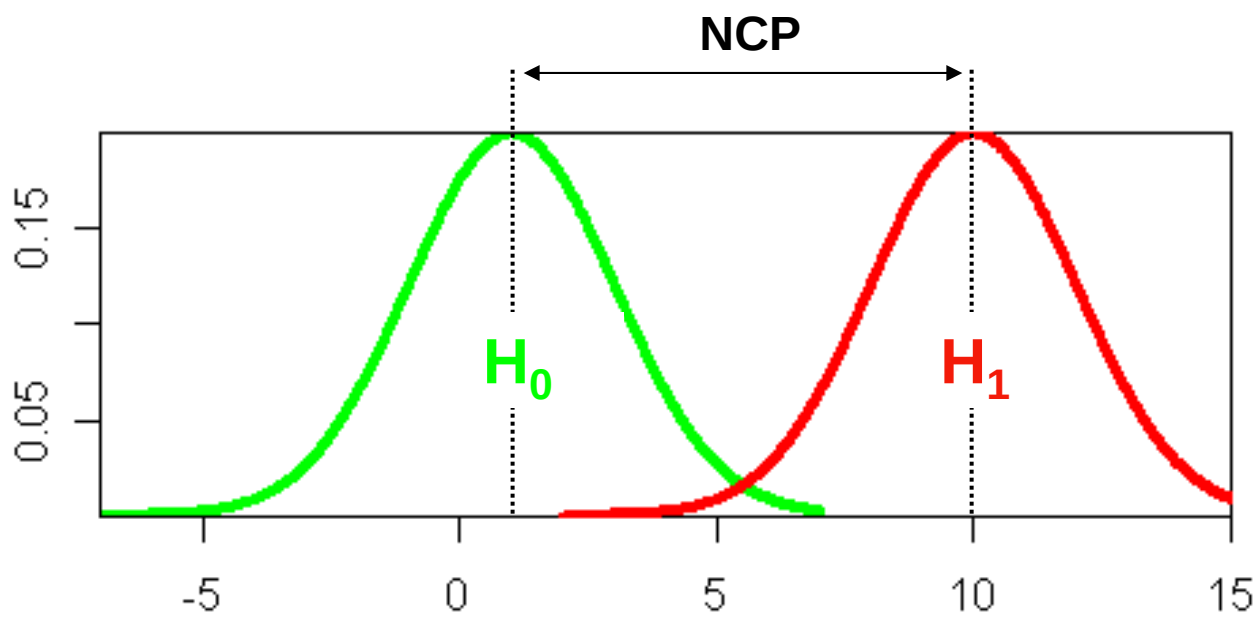
2. Try to shift the distribution of the linkage test statistic when H_1 is true as far as possible from the distribution when H_0 is true.

Non-centrality parameter



	Central X^2	Non-central X^2
Mean (μ)	**df**	**df + NCP**
Variance (σ^2)	**2*(df)**	**2*(df) + 4*NCP**

These distributions ARE NOT chi-sq with 1df!! Just for illustration.. Run R script in folder to see what they really look like..

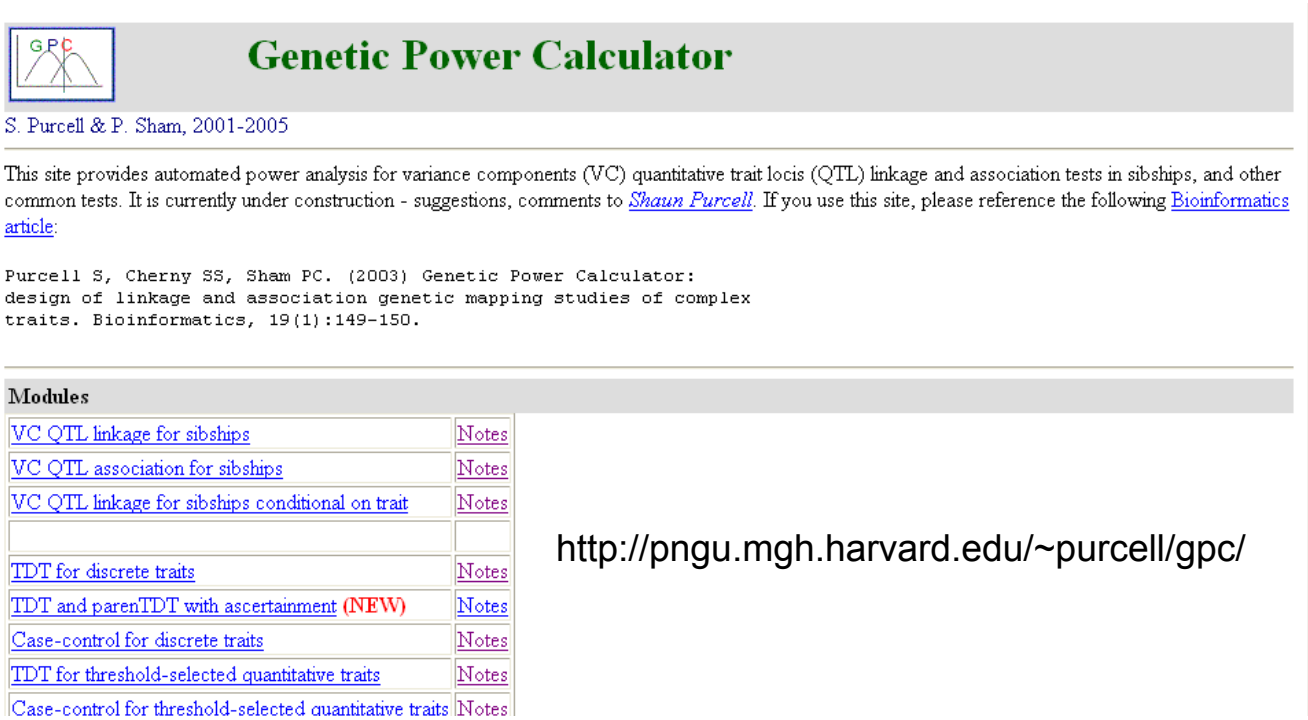


Small NCP → Big overlap between H_0 and H_1 distributions → **Lower power**

Large NCP → Small overlap between H_0 and H_1 distributions → **Greater power**

Short practical on GPC

- Genetic Power Calculator is an online resource for carrying out basic power calculations.



Genetic Power Calculator

S. Purcell & P. Sham, 2001-2005

This site provides automated power analysis for variance components (VC) quantitative trait locis (QTL) linkage and association tests in sibships, and other common tests. It is currently under construction - suggestions, comments to *Shaun Purcell*. If you use this site, please reference the following Bioinformatics article:

Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics, 19(1):149-150.

Modules

VC QTL linkage for sibships	Notes
VC QTL association for sibships	Notes
VC QTL linkage for sibships conditional on trait	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment **(NEW)**	Notes
Case-control for discrete traits	Notes
TDT for threshold-selected quantitative traits	Notes
Case-control for threshold-selected quantitative traits	Notes

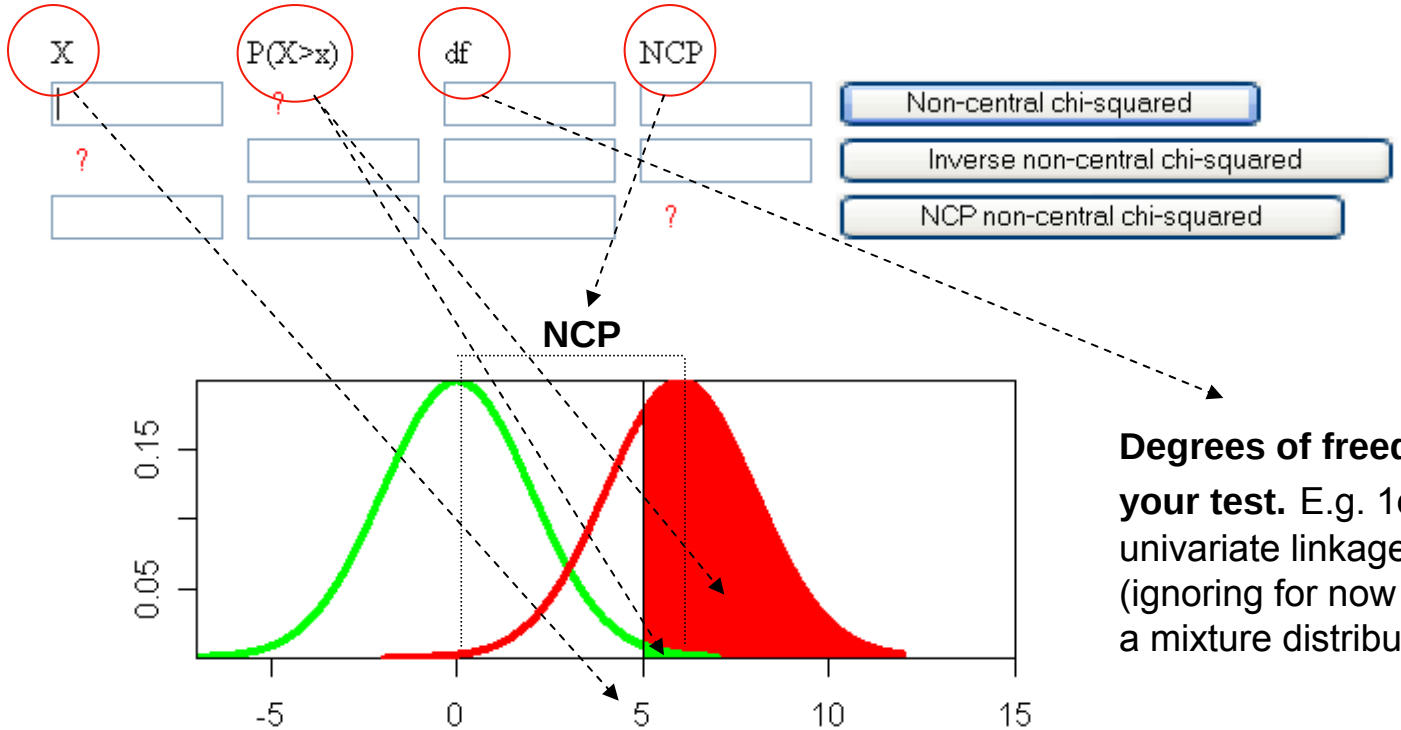
http://pngu.mgh.harvard.edu/~purcell/gpc/

- For our 1st example we will use the probability function calculator to play with power

Using the Probability Function Calculator of the GPC

1. Go to: ‘`http://pngu.mgh.harvard.edu/~purcell/gpc/`’
   Click the ‘Probability Function Calculator’ tab.

2. We’ll focus on the first 3 input lines. These refer to the chi-sq distribution that we’re interested in right now.



Degrees of freedom of your test. E.g. 1df for univariate linkage (ignoring for now that it’s a mixture distribution)

Exercises

1. Let's start with a simple exercise.

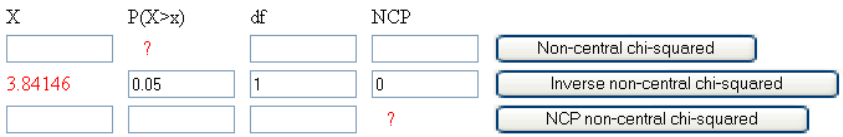
Determine the critical value (X) of a chi-square distribution with 1 df and NCP = 0, such that P(X>x) = 0.05.

df = 1

NCP = 0

P(X>x) = 0.05

X = ?



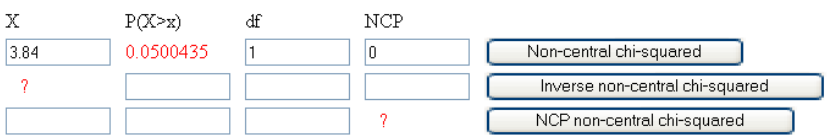
Determine the P(X>x) for a chi-square distribution with 1 df and NCP = 0 and X = 3.84.

df = 1

NCP = 0

P(X>x) = ?

X = 3.84



Exercises

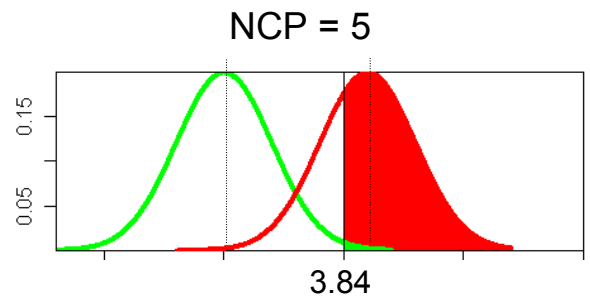
2. Find the power when the NCP of the test is 5, degrees of freedom=1, and the critical X is 3.84.

df = 1

NCP = 5

P(X>x) = ?

X = 3.84



X	P(X>x)	df	NCP	
3.84	0.608922	1	5	Non-central chi-squared
?				Inverse non-central chi-squared
			?	NCP non-central chi-squared

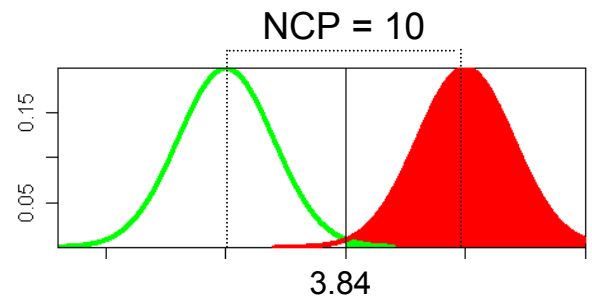
What if the NCP = 10?

df = 1

NCP = 10

P(X>x) = ?

X = 3.84



X	P(X>x)	df	NCP	
3.84	0.885451	1	10	Non-central chi-squared
?				Inverse non-central chi-squared
			?	NCP non-central chi-squared

Exercises

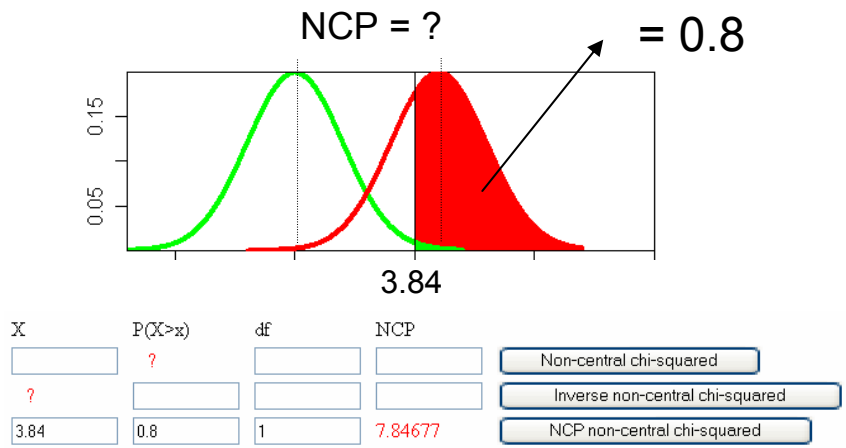
3. Find the required NCP to obtain a power of 0.8, for degrees of freedom=1 and critical X = 3.84.

df = 1

NCP = ?

P(X>x) = 0.8

X = 3.84



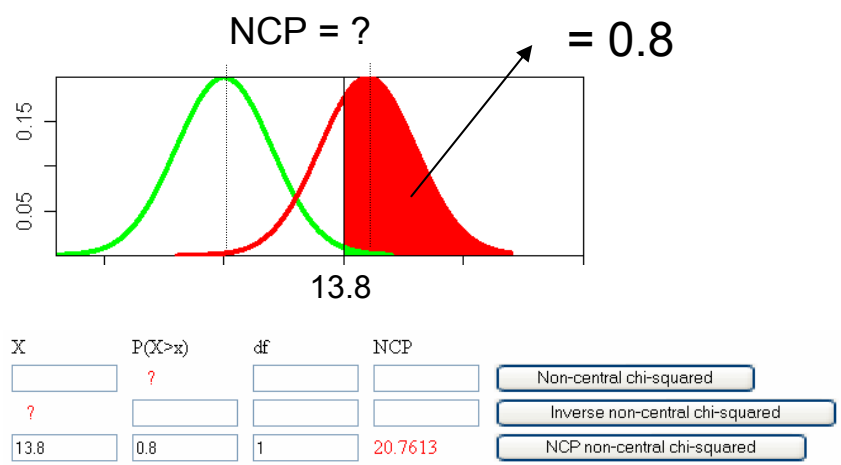
What if the X = 13.8?

df = 1

NCP = ?

P(X>x) = 0.8

X = 13.8



2. Estimate power for linkage and association

- Why is it important to estimate power?

  To determine whether the study you're designing/analysing can in fact localise the QTL you're looking for.

  Study design and interpretation of results.

  You'll need to do it for most grant applications.

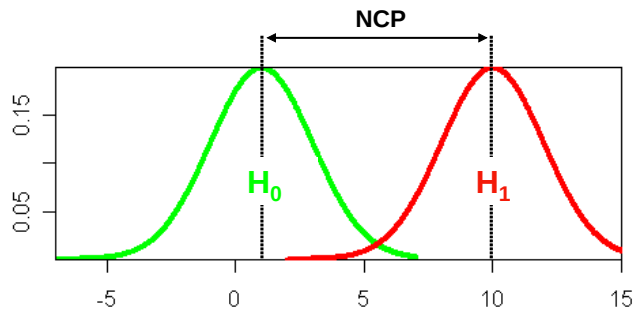
- When and how should I estimate power?

When?	How?
Study design stage	Theoretically, empirically
Analysis stage	Empirically

Theoretical power estimation

- NCP determines the power to detect linkage

  NCP = μ(H_1 is true) - df



- If we can predict what the NCP of the test will be, we can estimate the power of the test

Theoretical power estimation

Linkage

- Variance Components linkage analysis (and some HE extensions)

  Sham et al. 2000 *AJHG* 66: 1616

$$NCP \approx \frac{s(s-1)}{2} \frac{(1+r^2)}{(1-r^2)^2} \left[V_A^2 Var(\hat{\pi}) + V_D^2 Var(z) + V_A V_D Cov(\hat{\pi}, z) \right]$$

1. The number of sibs in the sibship (s)
2. Residual sib correlation (r)
3. Squared variance due to the additive QTL component (V_A)
4. Marker informativeness (i.e. Var($\hat{\pi}$) and Var(z))
5. Squared variance due to the dominance QTL component (V_D).

Another short practical on GPC

- The idea is to see how genetic parameters and the study design influence the NCP – and so the power – of linkage analysis

Using the 'VC QTL linkage for sibships' of the GPC

1. Go to: 'http://pngu.mgh.harvard.edu/~purcell/gpc/'
   Click the 'VC QTL linkage for sibships' tab.

Genetic Power Calculator

QTL Linkage for Sibships

QTL additive variance :
QTL dominance variance : No dominance (* see below)
Residual shared variance :
Residual nonshared variance :
Recombination fraction :

Sample Size :
Sibship Size : 2

User-defined type I error rate : 0.05 (0.00000001 - 0.5)
User-defined power: determine N : 0.80 (0 - 1)
(1 - type II error rate)

Process Reset

Exercises

1. Let's estimate the power of linkage for the following parameters:

QTL additive variance: 0.2

QTL dominance variance: 0

Residual shared variance: 0.4

Residual nonshared variance: 0.4

Recombination fraction: 0

Sample Size: 200

Sibship Size: 2

User-defined type I error rate: 0.05

User-defined power: determine N : 0.8

Power = 0.36 (alpha = 0.05)
Sample size for 80% power = 681 families

Exercises

2. We can now assess the impact of varying the QTL heritability

QTL additive variance: 0.4

QTL dominance variance: 0

Residual shared variance: 0.4

Residual nonshared variance: 0.4

Recombination fraction: 0

Sample Size: 200

Sibship Size: 2

User-defined type I error rate: 0.05

User-defined power: determine N : 0.8

Power = 0.73 (alpha = 0.05)
Sample size for 80% power = 237 families

Exercises

3. … the sibship size

QTL additive variance: 0.2

QTL dominance variance: 0

Residual shared variance: 0.4

Residual nonshared variance: 0.2

Recombination fraction: 0

Sample Size: 200

Sibship Size: 3

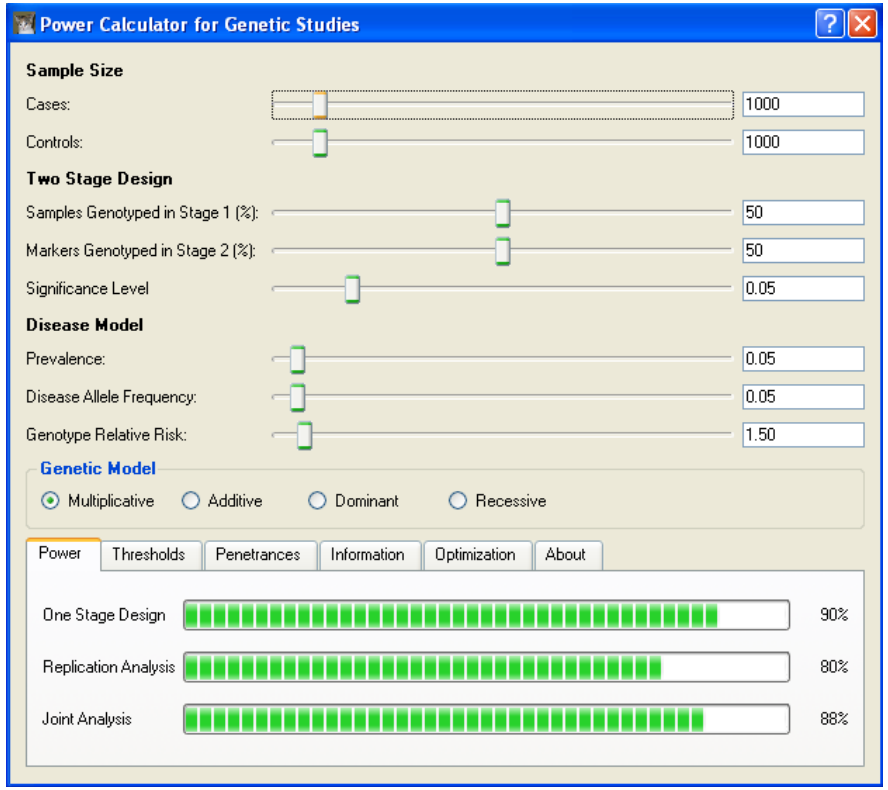
User-defined type I error rate: 0.05

User-defined power: determine N : 0.8

Power = 0.99 (alpha = 0.05)
Sample size for 80% power = 78 families

Theoretical power estimation *Association: case-control*

- CaTS performs power calculations for large genetic association studies, including two stage studies.



- http://www.sph.umich.edu/csg/abecasis/CaTS/index.html

Theoretical power estimation

Association: TDT

- TDT Power calculator, while accounting for the effects of untested loci and shared environmental factors that also contribute to disease risk

TDT Power calculator

Described in:
Ferreira, Sham, Daly & Purcell (2006) Family history of disease often decreases the power of family-based association studies (submitted).
See the reference section for a brief description of input paramters and output statistics.

Disease parameters

	k	Disease prevalence	[0.0001 - 0.9999]
	p	Allele frequency	[0.0001 - 0.9999]
	Vq	Total locus variance	[0 - 1]
	Va	Background additive genetic variance	[0 - 1]
	Vc	Shared environment variance	[0 - 1]
Add	Inheritance model		

Ascertainment parameters

	Number of families
1	N affected offspring per family required for selection
No	Ascertain discordant parents?

- http://pngu.mgh.harvard.edu/~mferreira/power_tdt/calculator.html

- Theoretical power estimation

Advantages: Fast, GPC, CaTS

Disadvantages: Approximation, may not fit well individual study designs, particularly if one needs to consider more complex pedigrees, missing data, ascertainment strategies, different tests, etc…

Empirical power estimation

- **Mx:** simulate covariance matrices for 3 groups (IBD 0, 1 and 2 pairs) according to an FQE model (i.e. with $V_Q > 0$) and then fit the wrong model (FE). The resulting test statistic (minus 1df) corresponds to the NCP of the test.

  See powerFEQ.mx script.

  Still has many of the disadvantages of the theoretical approach, but is a useful framework for general power estimations.

- **Simulate data:** generate a dataset with a simulated marker that explains a proportion of the phenotypic variance. Test the marker for linkage with the phenotype. Repeat this *N* times. For a given *α*, Power = proportion of replicates with a *P*-value < α (e.g. < 0.05).

Example with LINX

http://pngu.mgh.harvard.edu/~mferreira/

3. How to improve power

Factors that influence type-1 error and power

	Linkage	Association	
		Family-based	Case-control
1. Ascertainment Family structure, selective sampling	✓	✓	✓
2. Disease model QTL heritability, MAF, disease prevalence	✓	✓	✓
3. Deviations in trait distribution	✓	✓	
4. Pedigree errors	✓	✓	
5. Genotyping errors	✓	✓	✓
6. Missing data	✓	✓	✓
7. Genome coverage	✓	✓	✓

Pedigree errors

- Definition. When the self-reported familial relationship for a given pair of individuals differs from the real relationship (determined from genotyping data). Similar for gender mix-ups.

- Impact on linkage and FB association analysis. Increase type-1 error rate (can also decrease power)

- Detection. Can be detected using genome-wide patterns of allele sharing. Some errors are easy to detect. Software: GRR.

  Boehnke and Cox (1997), *AJHG* **61:**423-429; Broman and Weber (1998), *AJHG* 63:1563-4; McPeek and Sun (2000), *AJHG* 66:1076-94; Epstein et al. (2000), *AJHG* 67:1219-31.

- Correction. If problem cannot be resolved, delete problematic individuals (family)

Pedigree errors

Impact on linkage

- CSGA (1997) A genome-wide search for asthma susceptibility loci in ethnically diverse populations. *Nat Genet* **15:**389-92
- ~15 families with wrong relationships
- No significant evidence for linkage
- Error checking is essential!

Results

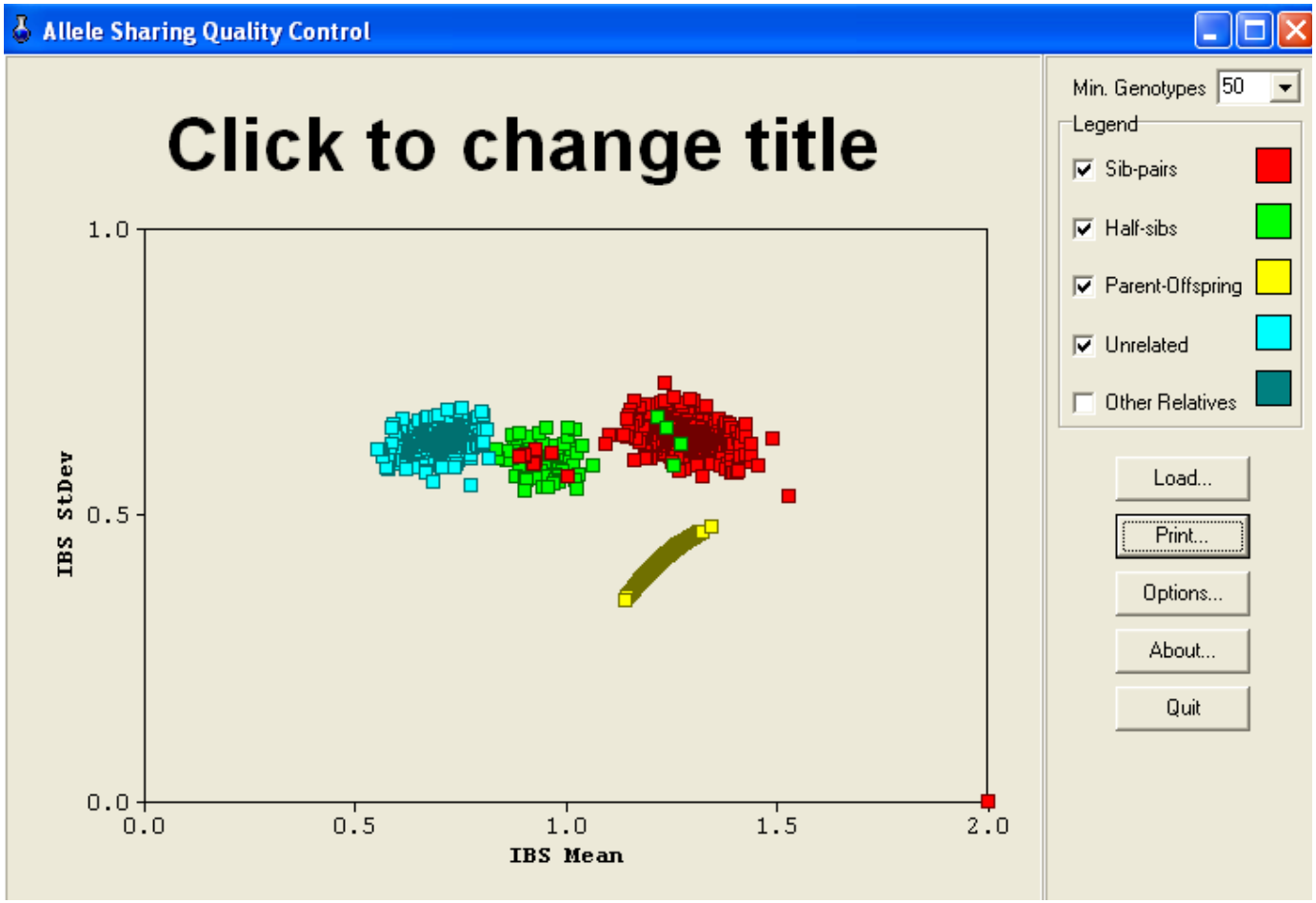
Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.

Genomewide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations

Margaret Gelder Ehm,[1] Maha C. Karnoub,[1] Hakan Sakul,[2,*] Kirby Gottschalk,[1] Donald C. Holt,[1] James L. Weber,[3] David Vaske,[3,†] David Briley,[1] Linda Briley,[1] Jan Kopf,[1] Patrick McMillen,[1] Quan Nguyen,[1] Melanie Reisman,[1] Eric H. Lai,[1] Geoff Joslyn,[2,‡] Nancy S. Shepherd,[1] Callum Bell,[2,§] Michael J. Wagner,[1] Daniel K. Burns,[1] and the American Diabetes Association GENNID Study Group[1]

Pedigree errors

Detection/Correction



GRR http://www.sph.umich.edu/csg/abecasis

Practical

▷ **Aim:** <u>Identify pedigree errors with GRR</u>

1. Go to: ‘`Egmondserver\share\Programs`’
   Copy entire ‘GRR’ folder into your desktop.

2. Go into the ‘GRR’ folder in your desktop, and run the GRR.exe file.

3. Press the ‘Load’ button, and navigate into the same ‘GRR’ folder on the desktop. Select the file ‘sample.ped’ and press ‘Open’. **Note that all sibpairs in ‘sample.ped’ were reported to be fullsibs or half-sibs.**

I’ll identify one error. Can you identify the other two?

Summmary

1. Statistical power
2. Estimate the power of linkage analysis
3. Improve the power of linkage analysis