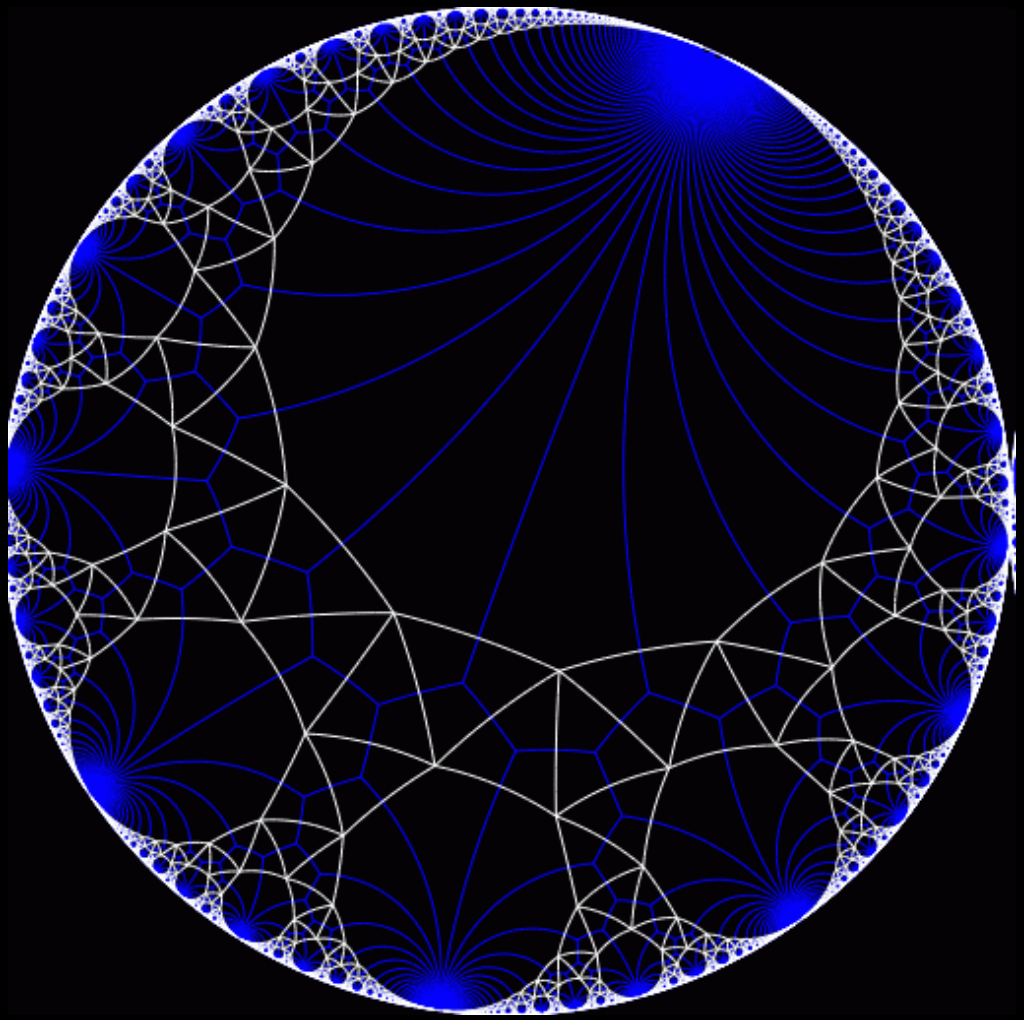
Multiple Testing, Permutation, False Discovery



Benjamin Neale
Pak Sham
Shaun Purcell

Hodgepodge anyone?

- Multiple testing
  - Where it comes from
  - Why is it a problem
- False discovery
  - Theory & practice
- Permutation
  - Theory & practice
- Additional handy techniques

Hodgepodge anyone?

- Multiple testing
  - Where it comes from
  - Why is it a problem
- False discovery
  - Theory & practice
- Permutation
  - Theory & practice
- Additional handy techniques

What do we test

- Raise your hand if:

What do we test

- Raise your hand if:
  - You have analyzed more than 1 phenotype on a dataset

What do we test

- Raise your hand if:
  - You have analyzed more than 1 phenotype on a dataset
  - Used more than one analytic technique on a dataset (e.g. single marker association and haplotype association)

What do we test

- Raise your hand if:
  - You have analyzed more than 1 phenotype on a dataset
  - Used more than one analytic technique on a dataset (e.g. single marker association and haplotype association)
  - Picked your best result from the bunch

Genome-wide association

High throughput genotyping

Other multiple testing considerations

- Genome-wide association is really bad
  - At 1 test per SNP for 500,000 SNPs
  - 25,000 expected to be significant at $p<0.05$, by chance alone

Other multiple testing considerations

- Genome-wide association is really bad
  - At 1 test per SNP for 500,000 SNPs
  - 25,000 expected to be significant at $p<0.05$, by chance alone
- To make things worse
  - Dominance (additive/dominant/recessive)
  - Epistasis (multiple combinations of SNPs)
  - Multiple phenotype definitions
  - Subgroup analyses
  - Multiple analytic methods

Bonferroni correction

- For testing 500,000 SNPs
  - 5,000 expected to be significant at $p<0.01$
  - 500 expected to be significant at $p<0.001$
  - ……
  - 0.05 expected to be significant at $p<0.0000001$
- Suggests setting significance level to $\alpha = 10^{-7}$*
- Bonferroni correction for m tests
  - set significance level for p-values to $\alpha = 0.05 / m$
  - (or adjust the p-values to $m \times p$, before applying the usual $\alpha = 0.05$ significance level)

- *See Risch and Merikangas 1999

Implication for sample size

Genetic Power Calculator

m	α	χ^2	NCP (80% power)	Ratio
1	0.05	3.84	7.85	1
500	10^{-4}	15.14	22.39	2.85
500×10^3	10^{-7}	28.37	38.05	4.85
500×10^6	10^{-10}	41.82	53.42	6.81

Large but not “impossible” increase in sample size

Technical objection

Conservative when tests are non-independent

- Nyholt (2004)
  - Spectral decomposition of correlation matrix
  - Effective number of independent tests
  - May be conservative: Salyakina et al (2005)
- False Discovery
- Permutation procedure

Philosophical objection

"Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference"
Perneger (1998)

- Counter-intuitive: interpretation of finding depends on the number of other tests performed
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of type 2 errors, i.e. of not rejecting the general null hypothesis when important effects exist

A Bayesian perspective

For each significant test, we can consider the probability that H_0 is in fact true (i.e. false positive probability)

Prob (H_0 True | H_0 Rejected)

Using Bayes' rule

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

False Discovery Rate

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Alpha level:
Rate of false positives

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Proportion of tests
that follow the null distribution

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Power to detect association

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

Proportion of tests that follow the alternative distribution

Taking the formula apart

$$\mathrm{P}(H_0 \mid p \le \alpha) = \frac{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(p \le \alpha \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(p \le \alpha \mid H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)}$$

False Discovery Rate

Alpha level:
Rate of false positives

Proportion of tests
that follow the null distribution

Power to detect
association

Proportion of tests
that follow the
alternative distribution

A Bayesian perspective

Re-expressing the equation in term of α:

$$\alpha = \frac{\mathrm{P}(H_0 \mid p \le \alpha)}{1 - \mathrm{P}(H_0 \mid p \le \alpha)} \frac{1-\pi_0}{\pi_0} \frac{1-\beta}{1}$$

A Bayesian perspective

Re-expressing the equation in term of α:

$$\alpha = \frac{\mathrm{P}(H_0 \mid p \le \alpha)}{1 - \mathrm{P}(H_0 \mid p \le \alpha)} \frac{1-\pi_0}{\pi_0} \frac{1-\beta}{1}$$

Power

False Discovery Rate

Proportion of tests that follows the null distribution

Implications

- Justification of traditional choice α=0.05
  - False positive rate ~ α, when π_0 ~ ½ and 1-$\beta \rightarrow 1$

Implications

- Justification of traditional choice α=0.05
  - False positive rate ~ α, when π_0 ~ ½ and 1-$\beta \rightarrow 1$
- Maintenance of low false positive rate requires α to be set proportional to
  - 1-β (power)
  - $(1-\pi_0)/\pi_0$ (proportion of tests that follow the null)

Implications

- Justification of traditional choice α=0.05
  - False positive rate ~ α, when π_0 ~ ½ and 1-$\beta \rightarrow 1$
- Maintenance of low false positive rate requires α to be set proportional to
  - 1-β (power)
  - $(1-\pi_0)/\pi_0$ (proportion of tests that follow the null)
- Multiple testing usually reflects lack of strong hypotheses and therefore associated with high π_0
  - Bonferroni adjustment effectively sets $\alpha \propto 1/m$, which is equivalent to assuming $\pi_0 = m/(m+1)$. But is this reasonable?

Fixed significance level

- Use fixed value of π_0 based on a guesstimate of the proportion of SNPs in the genome that have an effect, e.g. $1-\pi_0 = 25/10^7 = 2.5 \times 10^{-6}$
- Power = 0.8
- False positive rate = 0.05
- Then $\alpha \sim 10^{-7}$ (regardless of m)

Adaptive significance level

- Use the availability of multiple tests to our advantage, because the empirical distribution of p-values can inform us about the suitable significance level
- Suppose that out of 500,000 SNPs, 100 are observed to be significant at α=0.00001. Since the expected number of significant SNPs occurring by chance is 5, the false positive rate given by setting α=0.00001 is 5/100
- Therefore a desired false positive rate can be obtained by setting α appropriately, according to the observed distribution of p-values (False Discovery Rate method)

Hodgepodge anyone?

- Multiple testing
  - Where it comes from
  - Why is it a problem
- False discovery
  - Theory & practice
- Permutation
  - Theory & practice
- Additional handy techniques

Benjamini-Hochberg FDR method

Benjamini & Hochberg (1995) Procedure:

1. Set FDR (e.g. to 0.05)
2. Rank the tests in ascending order of p-value, giving $p_1 \leq p_2 \leq \ldots \leq p_r \leq \ldots \leq p_m$
3. Then find the test with the highest rank, r, for which the p-value, p_r, is less than or equal to $(r/m) \times$ FDR
4. Declare the tests of rank 1, 2, …, r as significant

A minor modification is to replace m by $m\pi_0$

B & H FDR method

FDR=0.05

Rank	P-value	(Rank/m)×FDR	Reject H_0 ?
1	.008	.005	1
2	.009	.010	1
3	.165	.015	0
4	.205	.020	0
5	.396	.025	0
6	.450	.030	0
7	.641	.035	0
8	.781	.040	0
9	.901	.045	0
10	.953	.050	0

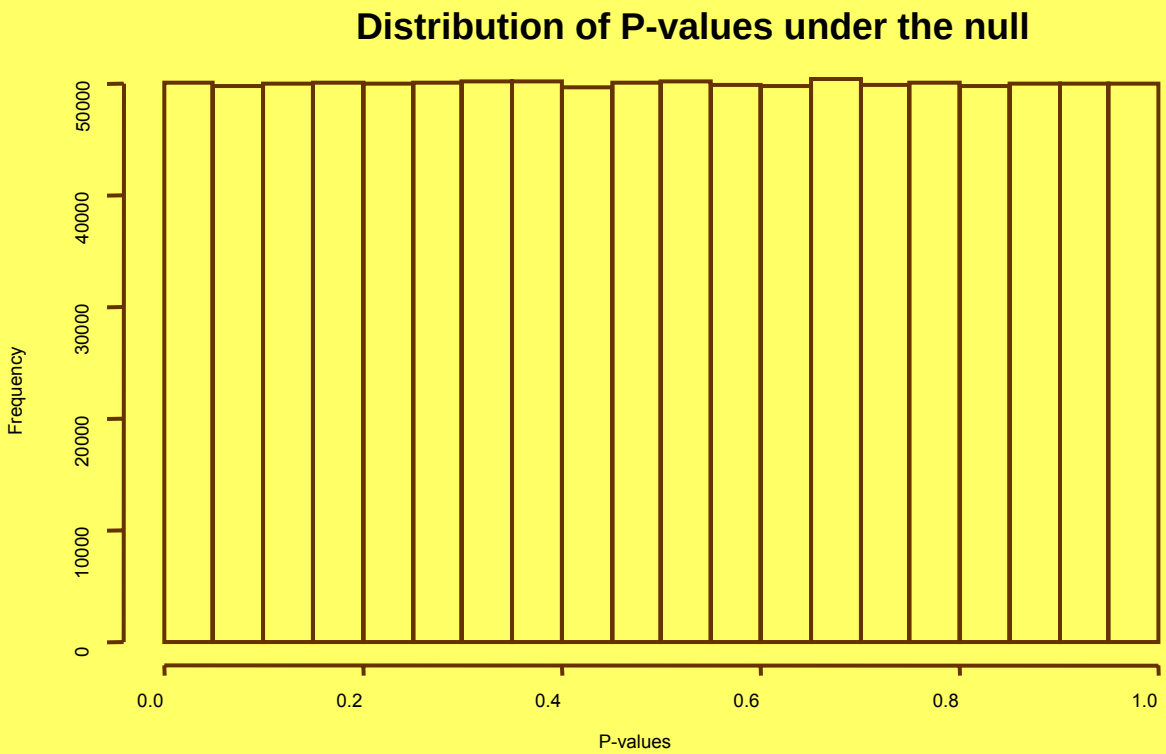
Practical example

- Excel worksheet, fdr.xls in \\faculty\ben
- Download to your directory
- 850 tests, with P-values
- FDR procedure in Excel
- Play around with changing the FDR level to see the behaviour of accepting/rejecting
- To determine which tests are accepted:
  - Start at the bottom (lowest rank)
  - Work up the list to find the 1st accept
  - That 1st accept and all tests above are accepted

Modified FDR methods

Storey 2002 procedure:

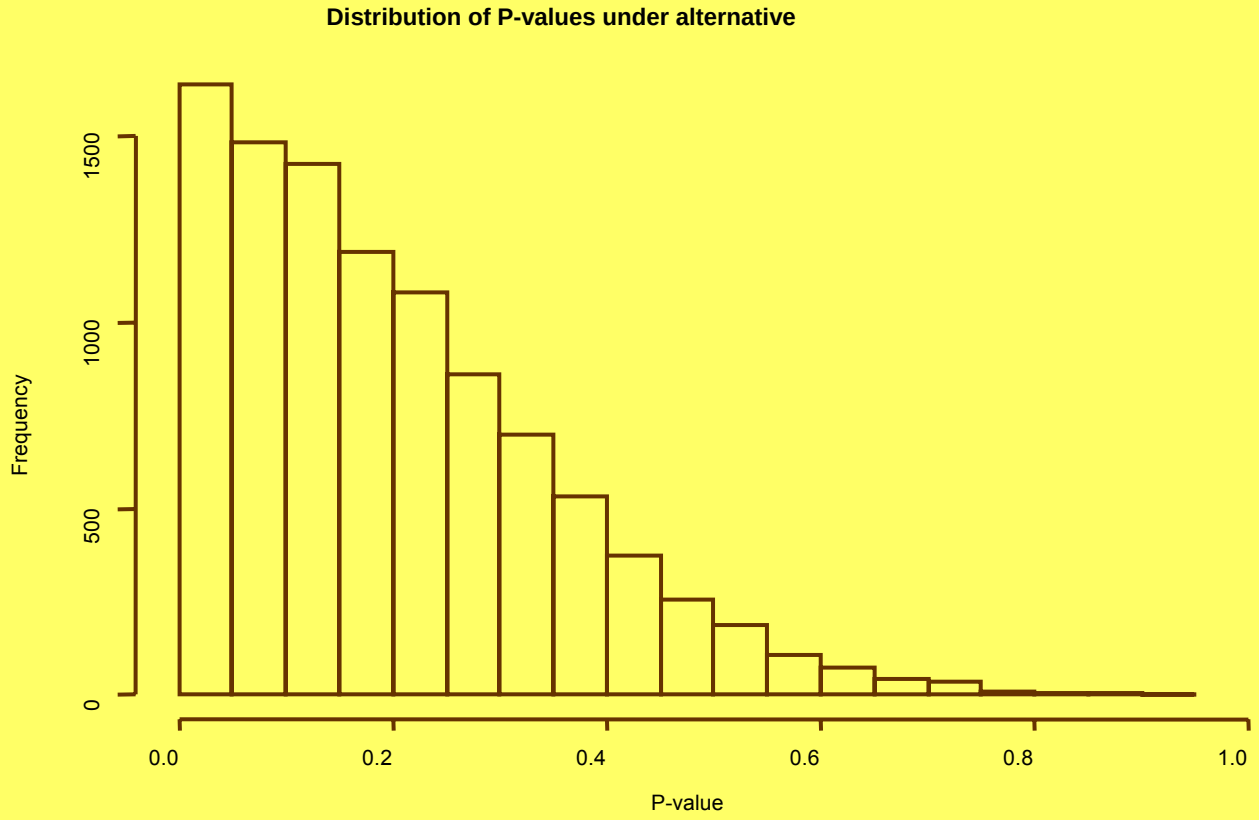
Under the null P-values look like:



Modified FDR methods

Storey 2002 procedure:

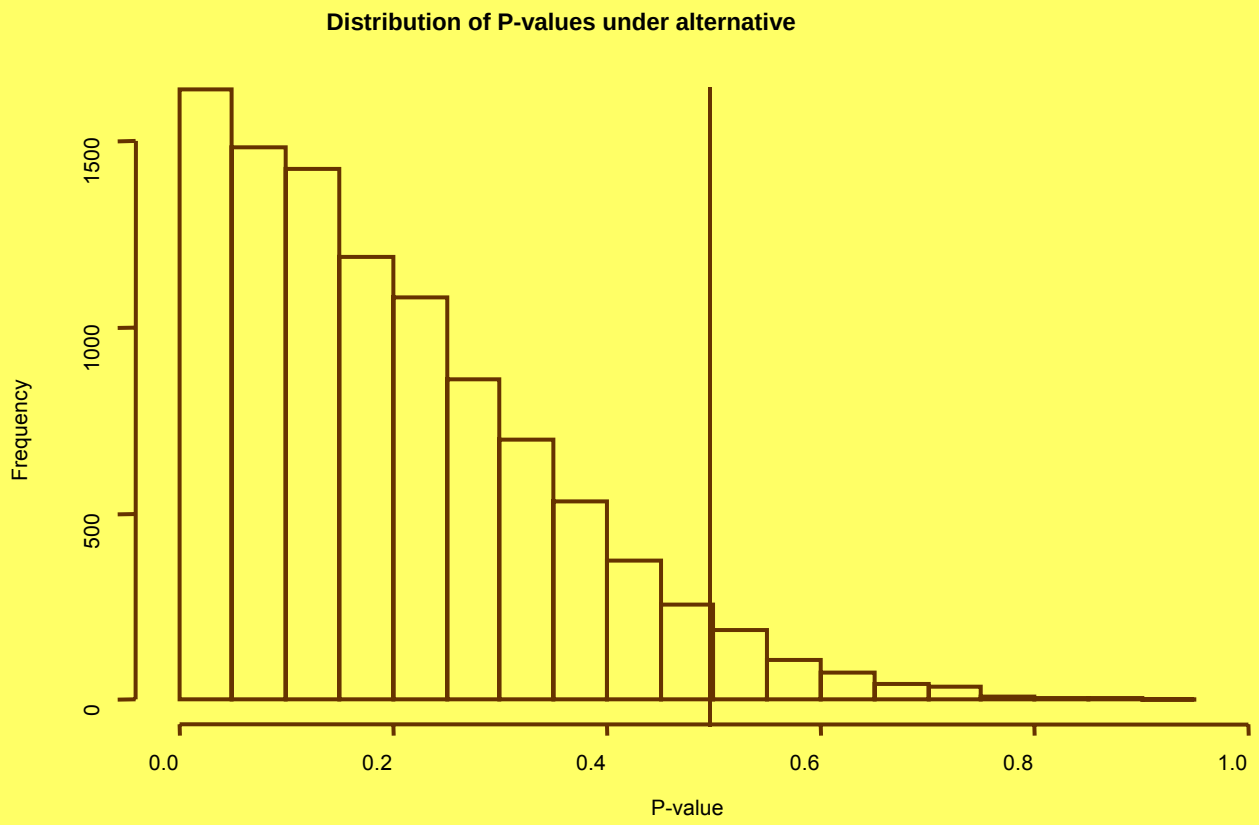
Under the alternative P-values look like:



Modified FDR methods

Storey 2002 procedure:

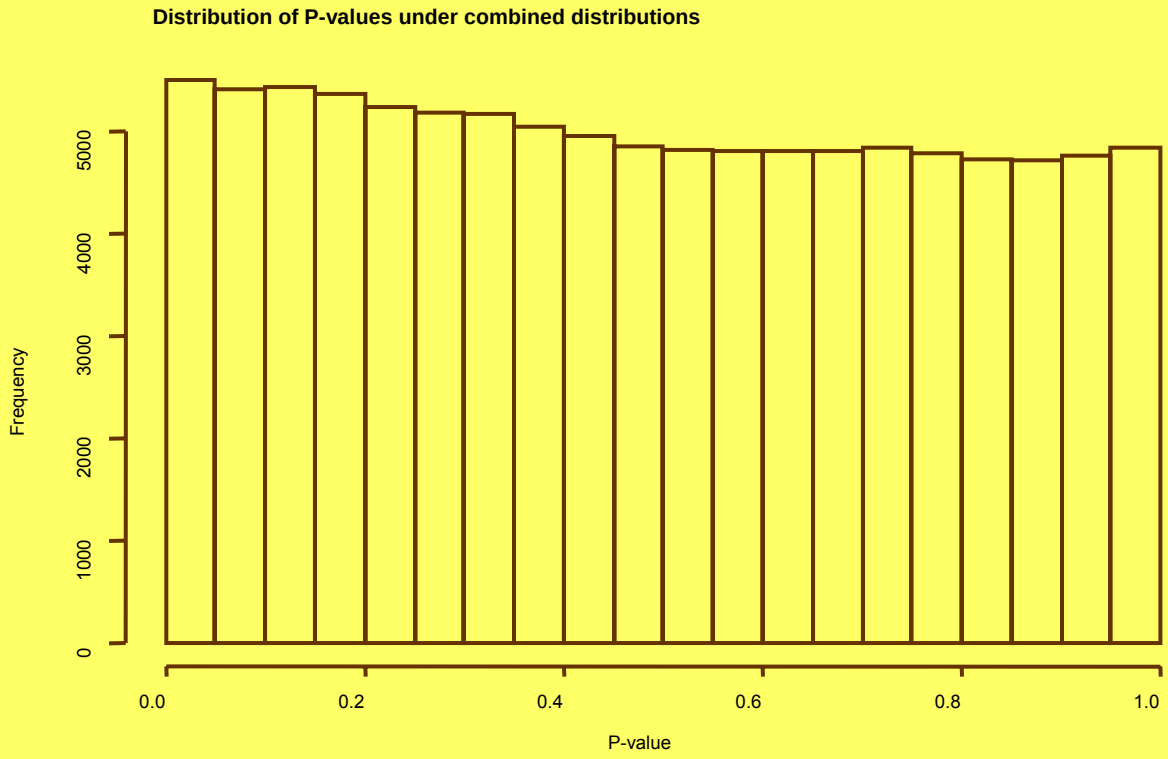
Under the alternative P-values look like:



Modified FDR methods

Storey 2002 procedure:

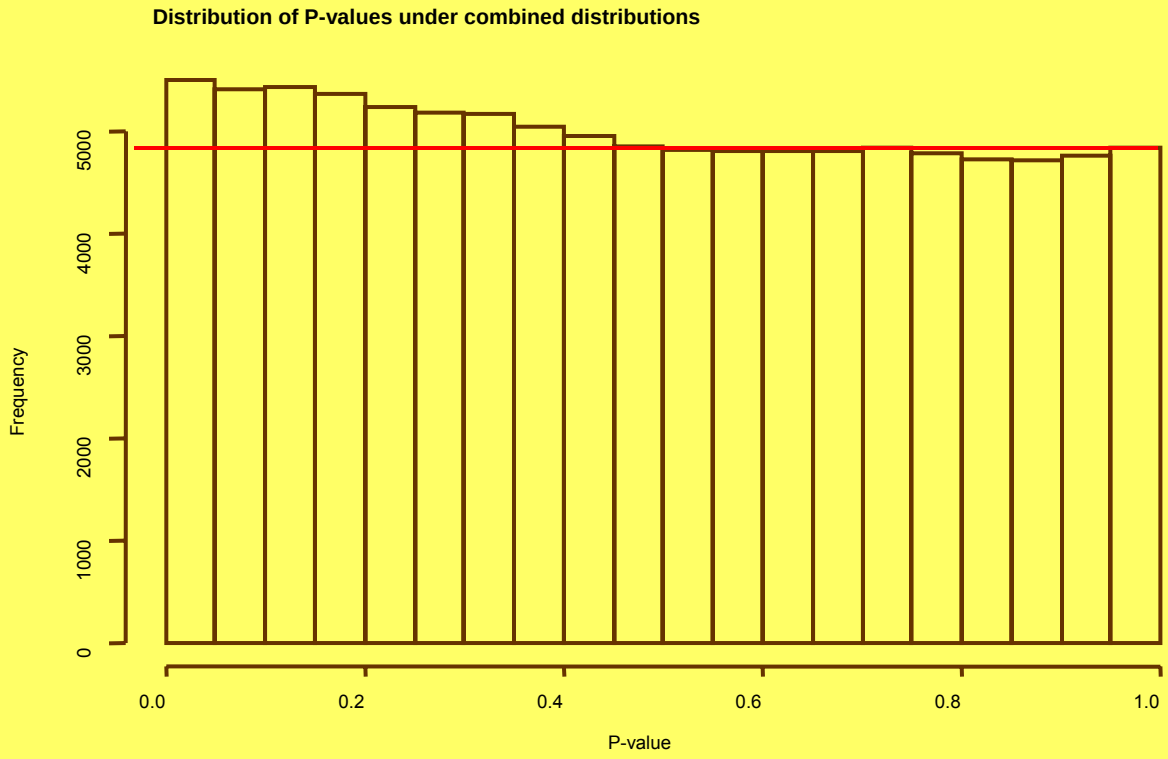
Combined distribution of P-values look like:



Modified FDR methods

Storey 2002 procedure:

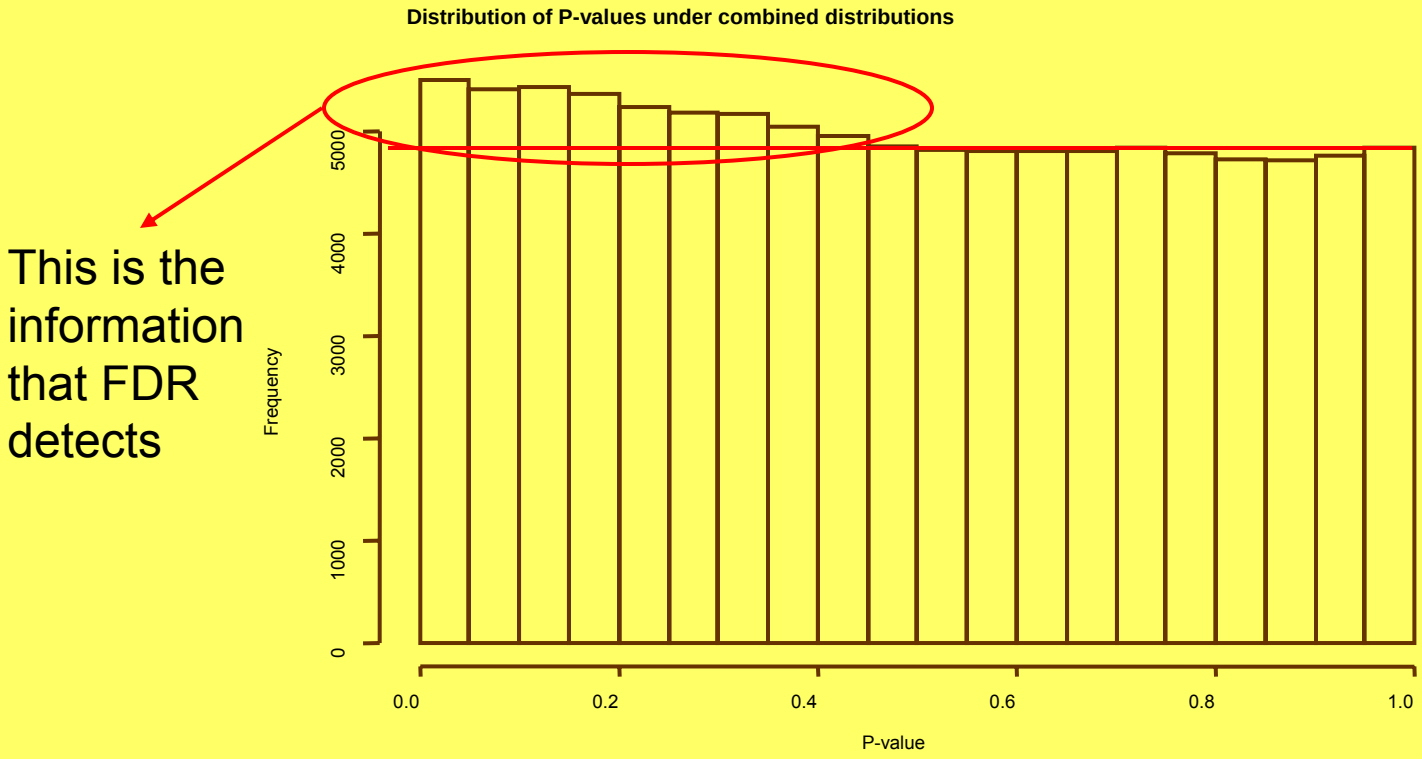
Combined distribution of P-values look like:



Modified FDR methods

Storey 2002 procedure:

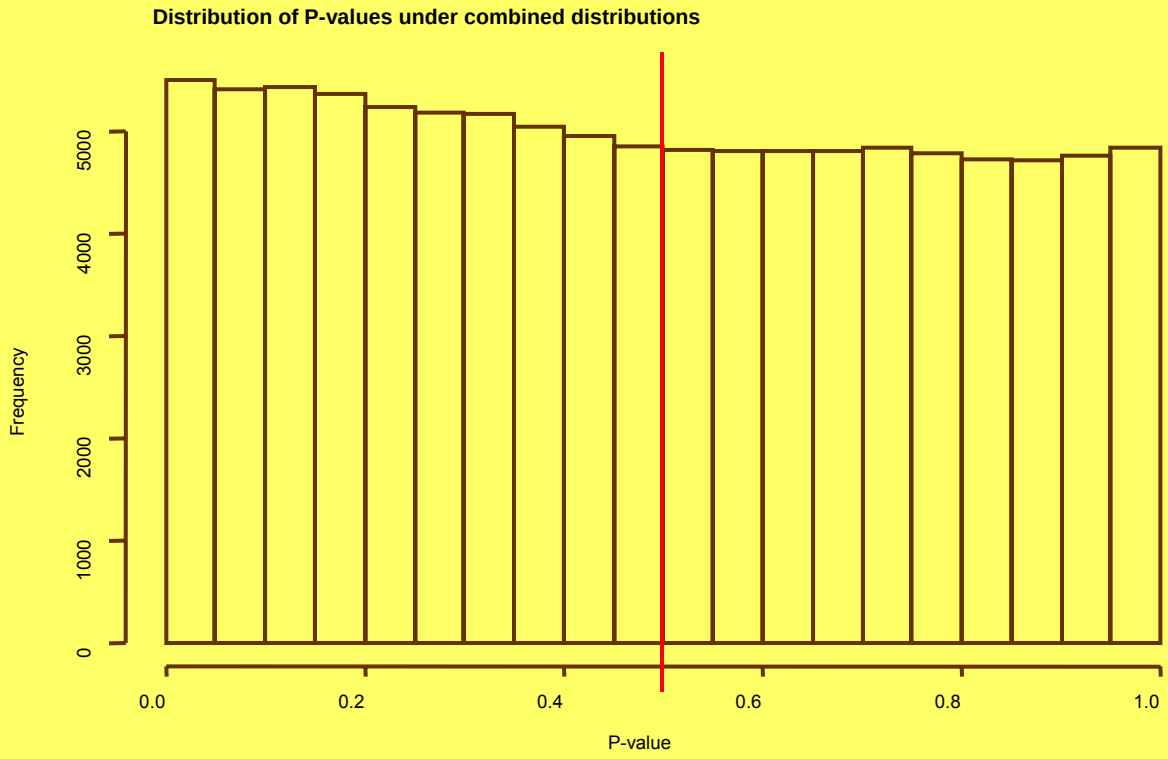
Combined distribution of P-values look like:



Modified FDR methods

Storey 2002 procedure:

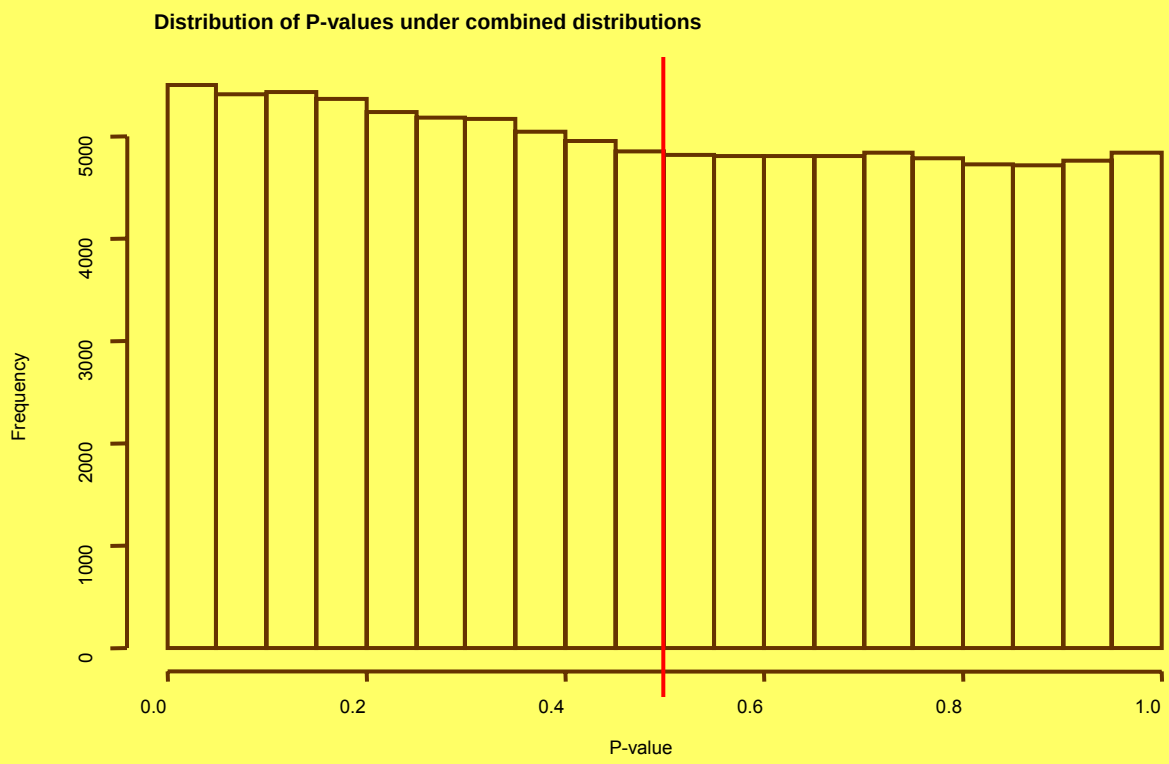
Combined distribution of P-values look like:



Modified FDR methods

Storey 2002 procedure:

Combined distribution of P-values look like:

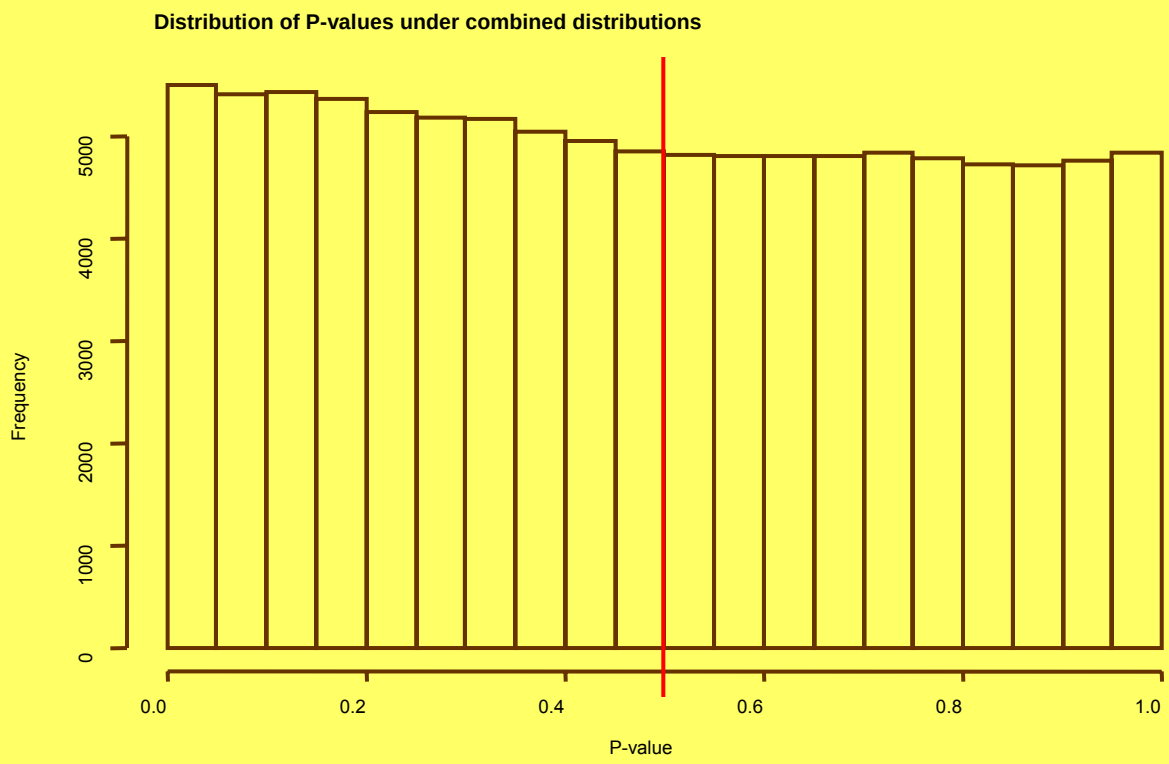


The number of tests above p = .5 is 47651out of 100000

Modified FDR methods

Storey 2002 procedure:

Combined distribution of P-values look like:



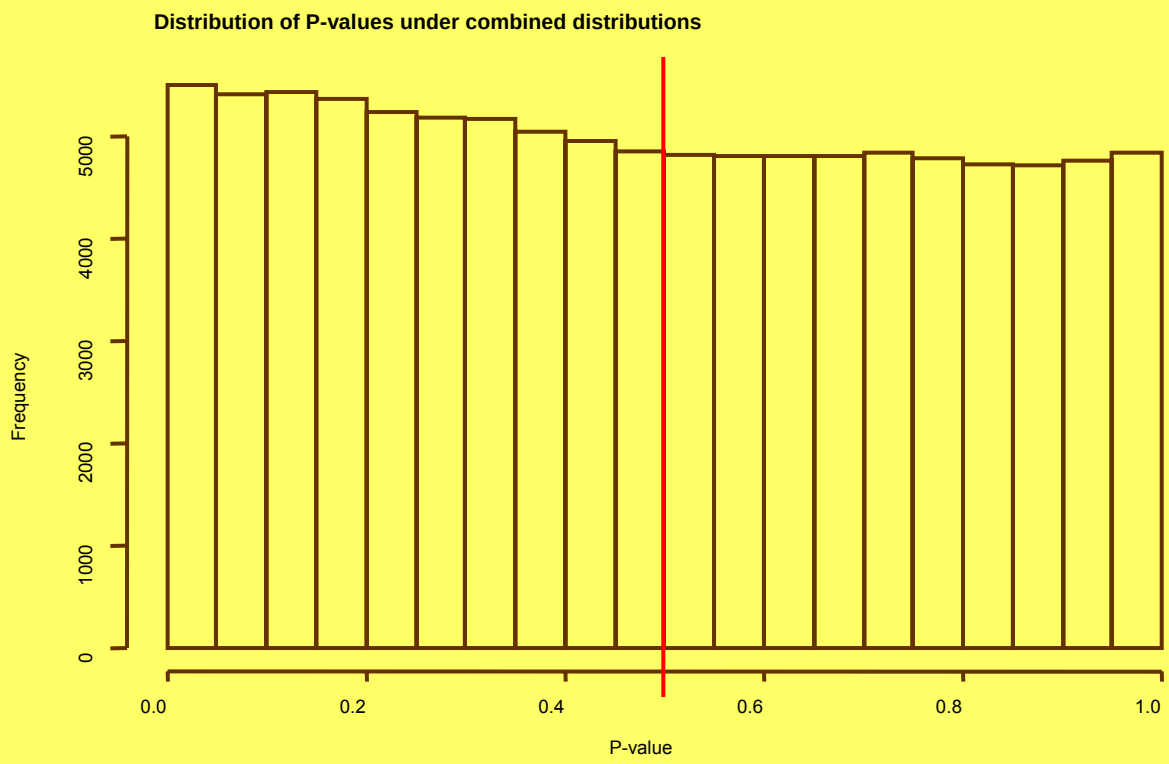
The number of tests above p = .5 is 47651out of 100000

So the proportion of tests that follows the null: 47651/50000 or .95302 = π_0

Modified FDR methods

Storey 2002 procedure:

Combined distribution of P-values look like:



The number of tests above p = .5 is 47651out of 100000

So the proportion of tests that follows the null: 47651/50000 or .95302 = π_0

So we replace the number of tests with the number of tests times π_0 or 95302.

“Parametric FDR” methods

Mixture model: some test statistics follow the null distribution, while others follow a **specified** alternative distribution

Special cases:

- Central and non-central chi-square distributions (Everitt & Bullmore, 1999)
- Central and non-central normal distributions (Cox & Wong, 2004)
- Uniform and beta distributions (Allison et al, 2002)

From fitted model, calculates the posterior probability of each test belonging to the null distribution (i.e. of being a false discovery if declared significant)

Pitfalls of the FDR method

- Assumption: p-values are distributed as U[0,1] under H_0
  - If untrue (e.g. biased genotyping, population substructure) then this could lead to an excess of small p-values and hence misleading FDR results
- Requires a large number of tests to work
- The accuracy of the FDR is not easy to determine
- Requires a distribution (detectable number) of tests under the alternative

Who came up with permutation?

- Hint: it's a statistical tool
- R. A. Fisher
- Proposed as validation for Student's t-test in 1935 in Fisher's The Design of Experiments
- Originally included all possible permutations of the data

Basic Principle

1. Under the null, all data comes from the same distribution
2. We calculate our statistic, such as mean difference
3. We then shuffle the data with respect to group and recalculate the statistic (mean difference)
4. Repeat step 3 multiple times
5. Find out where our statistic lies in comparison to the null distribution

Real Example

- Case-Control data, and we want to find out if there is a mean difference

	case		control
1	-0.49274	10	1.471227
2	-0.30228	11	0.612679
3	0.093007	12	-0.47886
4	0.715722	13	0.746045
5	1.272872	14	0.871994
6	-1.37599	15	0.985237
7	-0.14798	16	-0.44421
8	-1.22195	17	0.246393
9	1.2812	18	0.68246
Mean	-0.01979		0.52144

Mean difference .541

Permutation One

Note how the different labels have been swapped for the permutation

	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Mean difference = .329

Permutation One

Note how the different labels have been swapped for the permutation

Repeat many many many many times (and then repeat many more times)

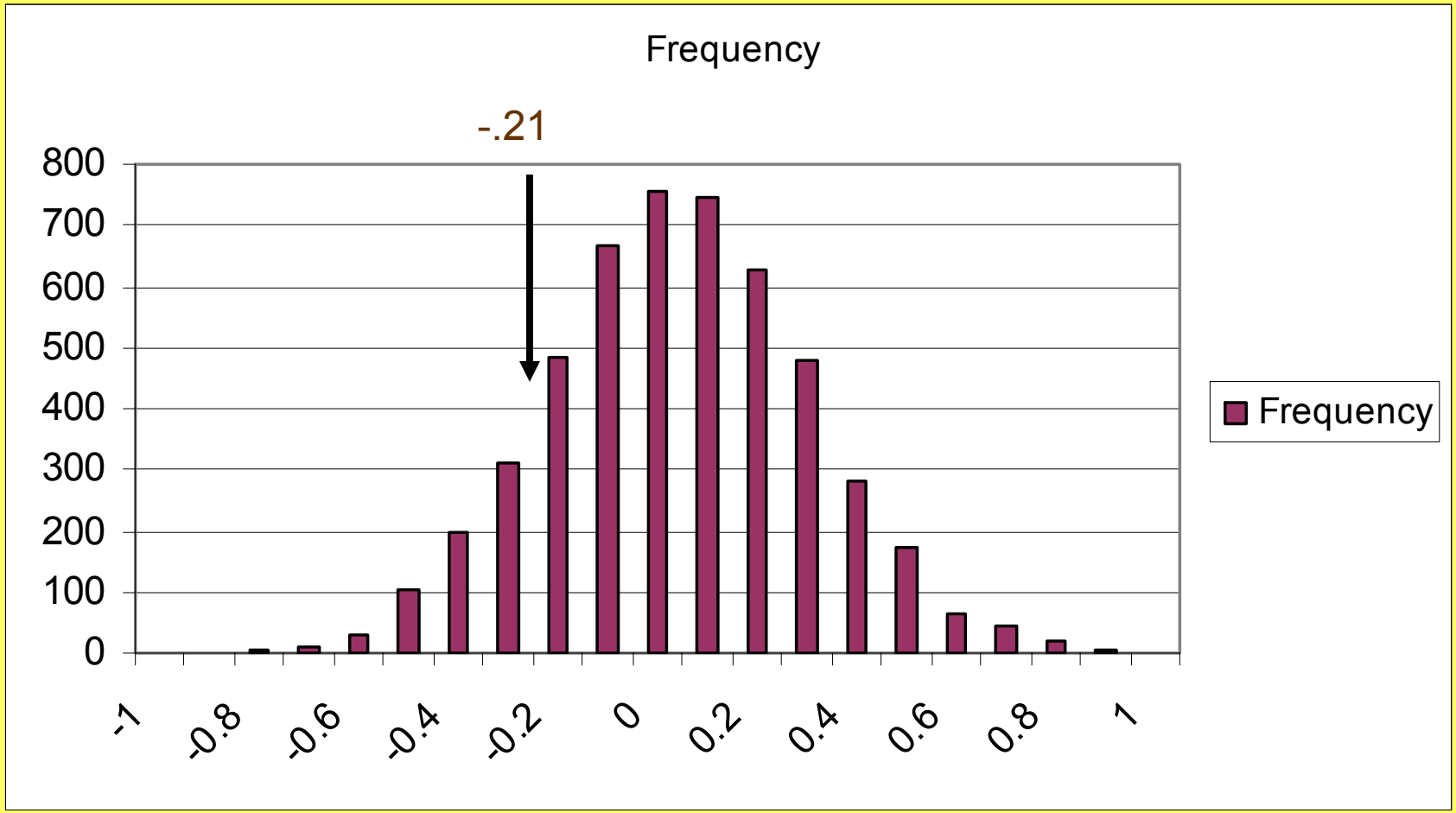
	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Mean difference = .329

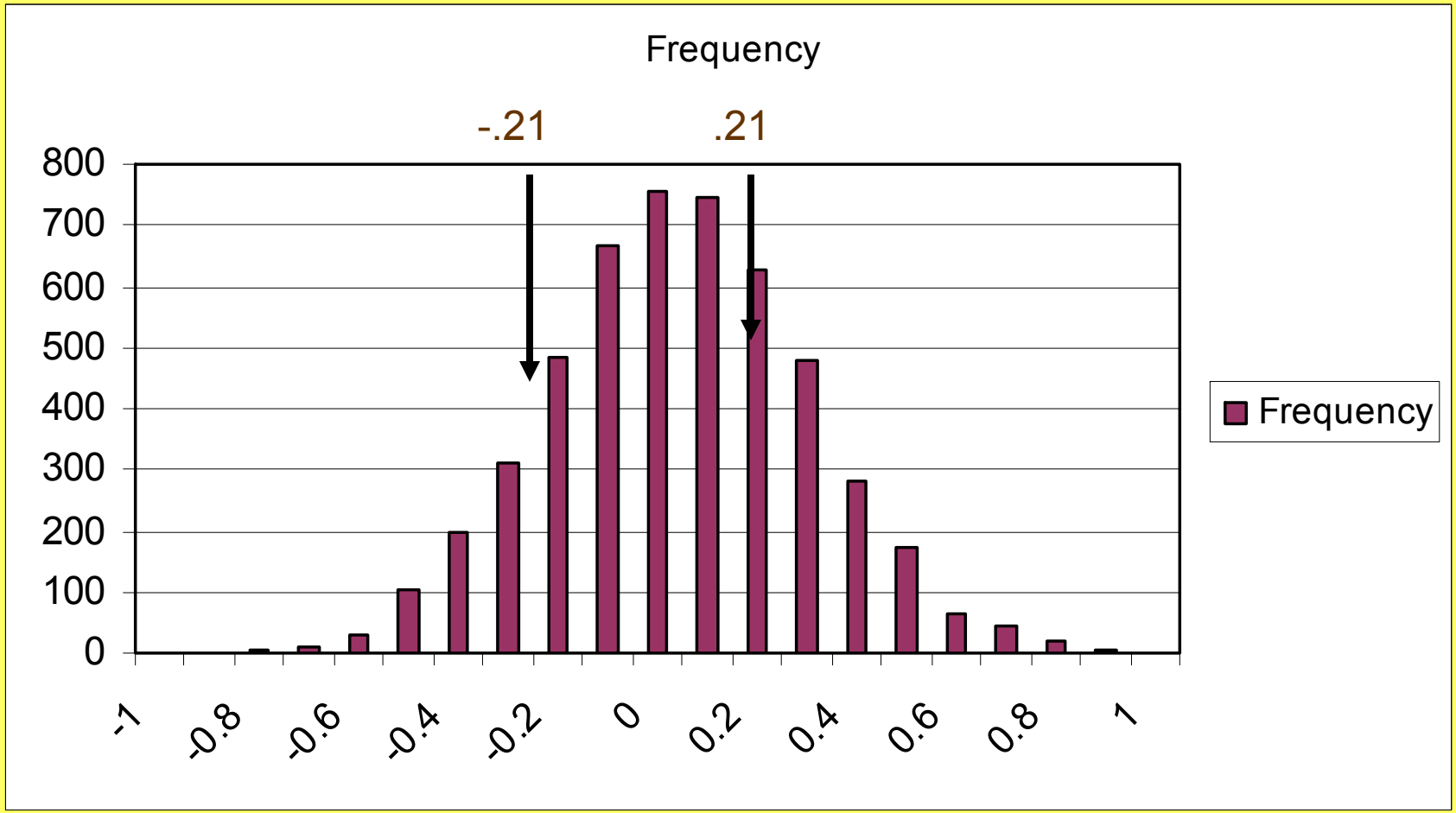
Simulation example

- I simulated 70 data points from a single distribution—35 cases and 35 controls
- Mean difference of -.21
- I then permuted randomly assigning case or control status
- Empirical significance= (#hits+1)/(#permutations+1)

Distribution of mean differences from permutations



Distribution of mean differences from permutations



Empirical Significance

- #hits is any permuted dataset that had a mean difference >.21 or <-.21
- #permutations is the trials permuted datasets we generate
- Result(#hits + 1/#permutations + 1) = 2025/5001 = .4049
- T test results = .3672

General advantages

- Does not rely on distributional assumptions
- Corrects for hidden selection
- Corrects for hidden correlation

General principles of permutation

- Disrupt the relationship being tested
  - Mean difference between group: switch groups
  - Test for linkage in siblings: randomly reassign the ibd sharing
  - If matched control then within pair permute

Practical example for permutation

- Permutation routine:
  - Case Control analysis
- Use R for permutation
- Many genetic programs have in-built permutation routines

R permutation

- R has an in-built simulate P-value function for chi square
- We'll start with that and progress to manual permutation to understand the process
- In \\workshop\faculty\ben
  - Copy both rscript.R, casecontrol.txt, and chiextract

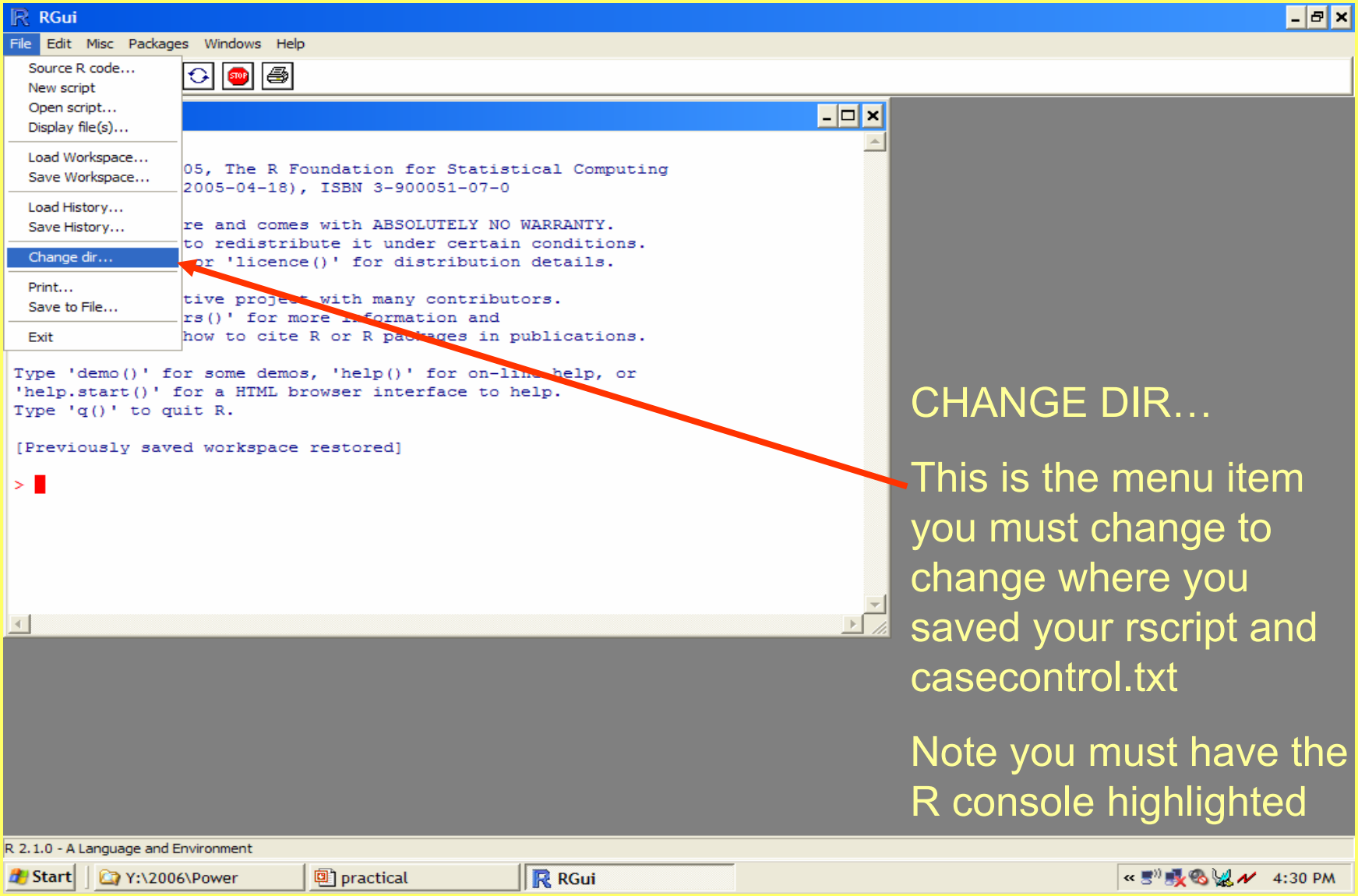
File descriptions

- rprog.R
  - Contains the script that generates the R simulated P and the manual simulations
- casecontrol.txt
  - Contains the data for the true chi square

Running the script

- Save all files to your directory in a convenient folder
- Fire up R
- Change the working directory in R to where the script and data are
  - To do this click on file menu then change working directory to either browse or type in the directory name
- In R type or follow the dialogues source(“rscripR”)
- That runs the script and some output will be generated and reported back from R

Picture of the menu

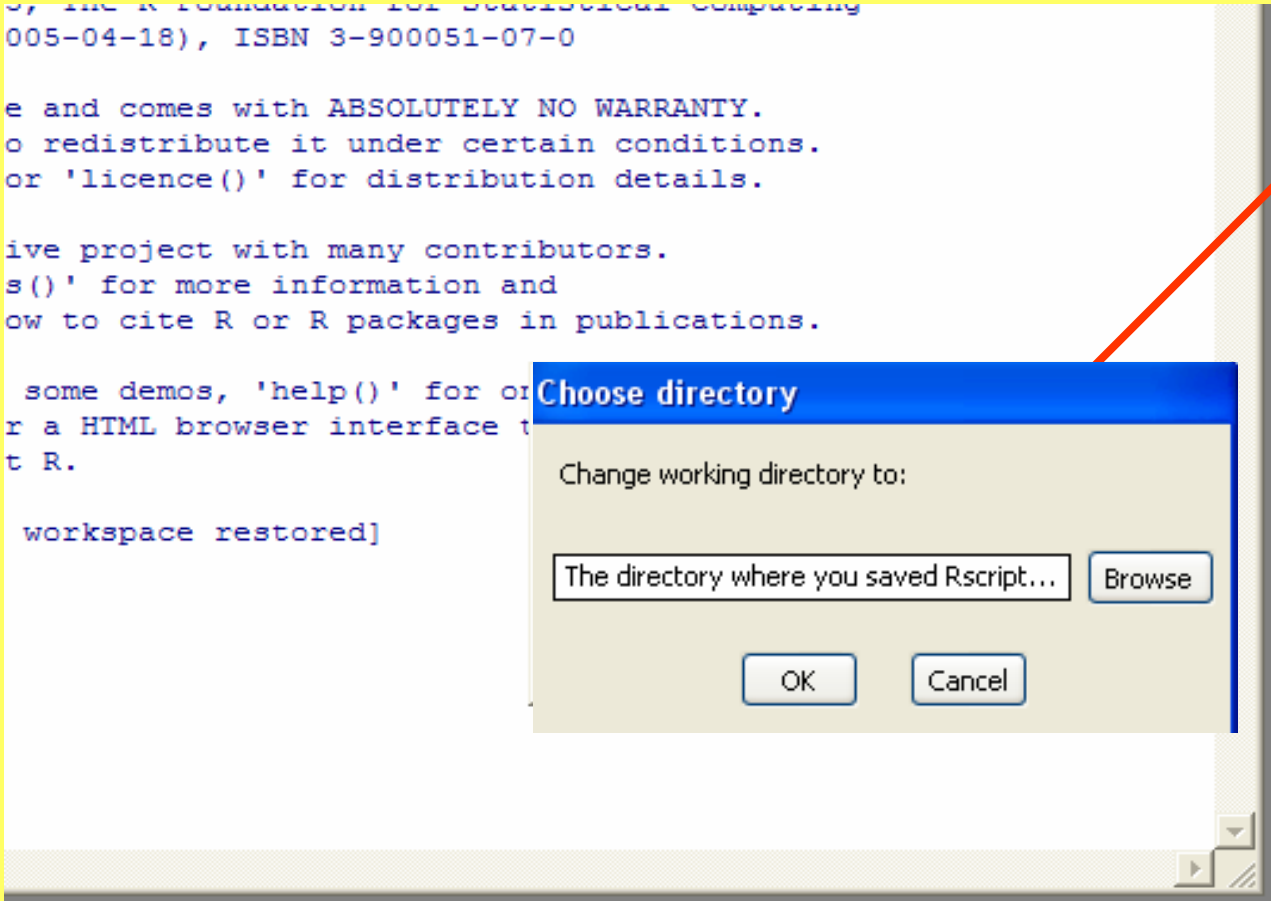


CHANGE DIR…

This is the menu item you must change to change where you saved your rscript and casecontrol.txt

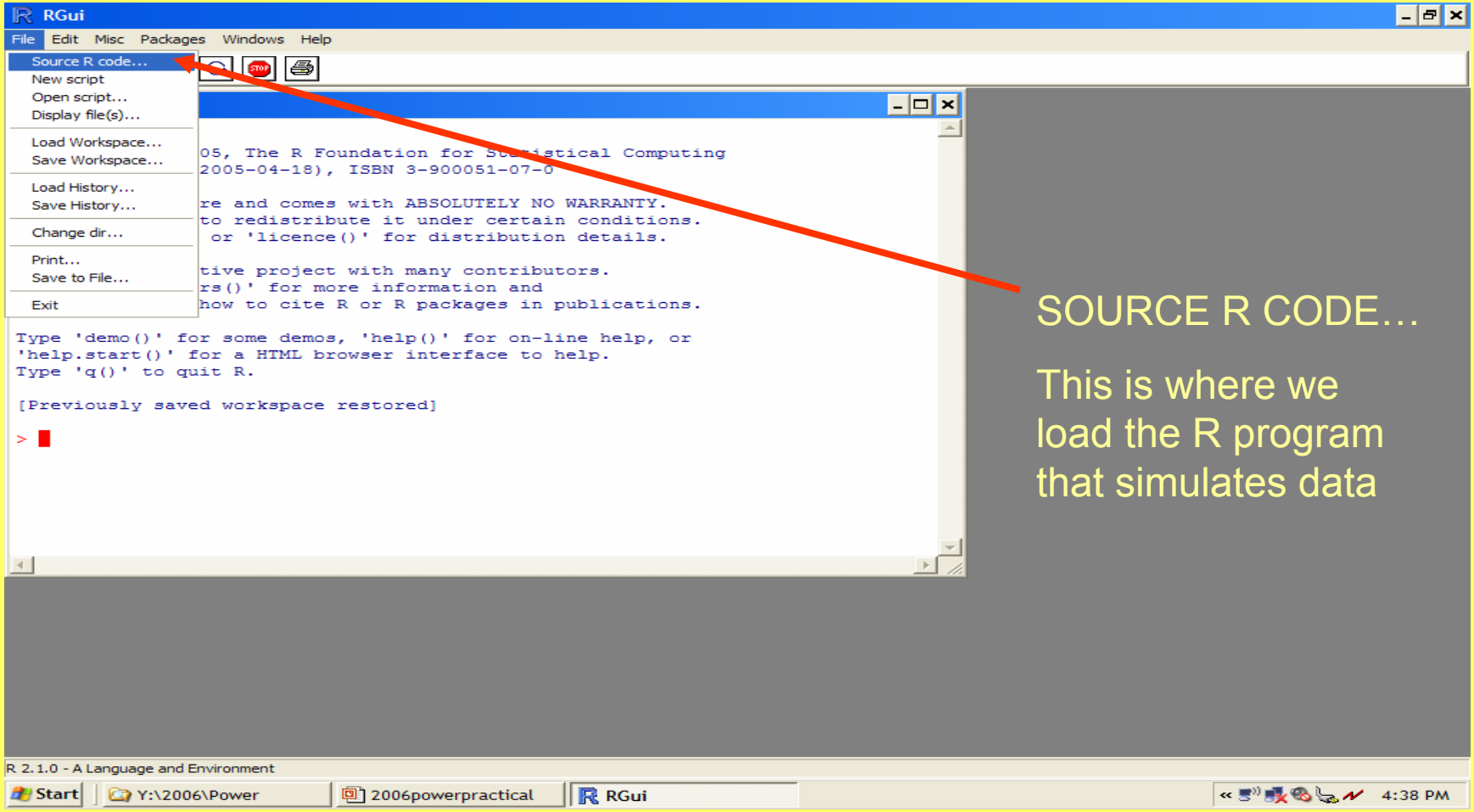
Note you must have the R console highlighted

Picture of the dialog box



Either type the path name or browse to where you saved the R script

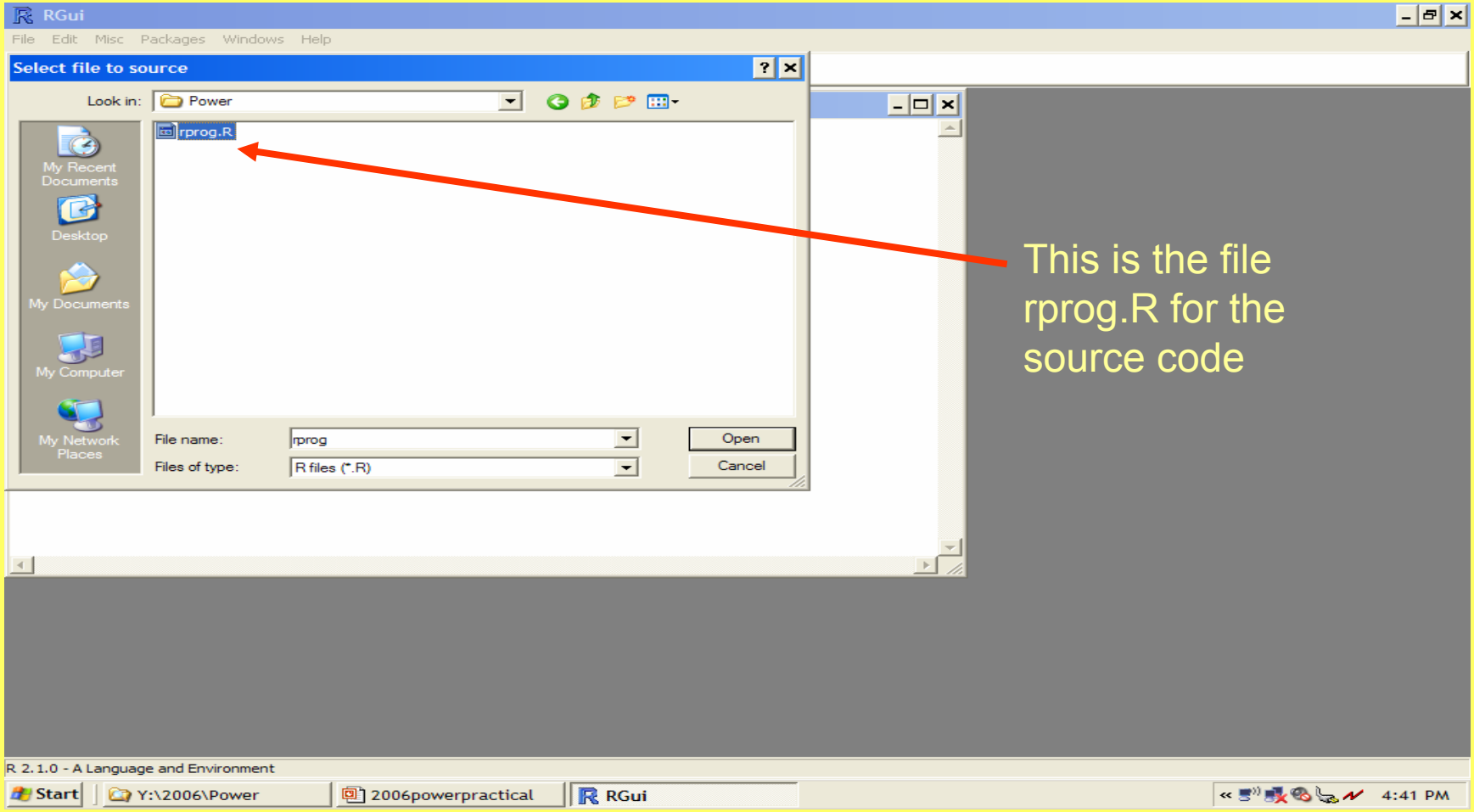
Running the R script



SOURCE R CODE…

This is where we load the R program that simulates data

Screenshot of source code selection



How would we do QTL permutation in Mx?

1. We analyze our real data and record χ^2
2. For sibpairs we shuffle the ibd probabilities for each sibpair
3. We reanalyze the data and record the new χ^2
4. We generate a distribution of χ^2 for the permuted sets
5. Place our statistic on the distribution
6. Repeat for all locations in genome

Some caveats

- Computational time can be a challenge
- Determining what to maintain and what to permute
- Variable pedigrees also pose some difficulties
- Need sufficient data for combinations
- Unnecessary when no bias, but no cost to doing it
- Moderators and interactions can be tricky